BEHAVIOR ANALYSIS

An Operant Analysis of

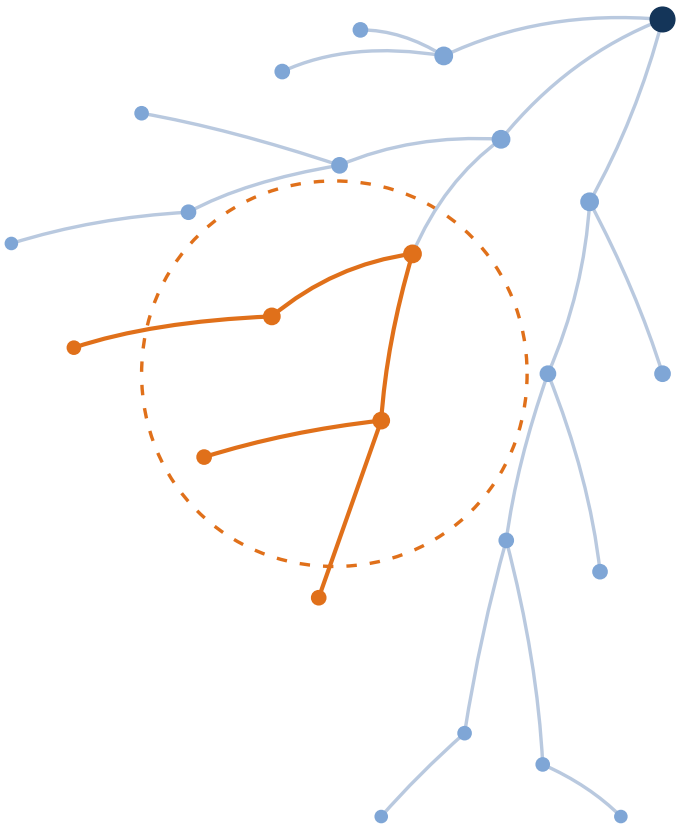
Hate Speech and Disinformation



Márcio Moreira · Carlos Augusto de Medeiros · Cristiano Coelho
Diogo Conque Seco-Ferreira · Frederico Neves Condé · Lucas Ferraz Córdova

iW4 instituto WALDEN4
walden4.com.br

instituto WALDEN4



BEHAVIOR ANALYSIS

An Operant Analysis of

Hate Speech and Disinformation

Márcio Moreira

Cristiano Coelho

Frederico Neves Condé

Carlos Augusto de Medeiros

Diogo Conque Seco-Ferreira

Lucas Ferraz Córdova

Instituto Walden4

Brasília, DF · 2026

CATALOGING RECORD

DDC-150

International Cataloging-in-Publication Data
(Prepared by the publisher)

An operant analysis of hate speech and disinformation [electronic book] / Márcio Borges Moreira ... [et al.] ; translated from the Brazilian Portuguese by Márcio Borges Moreira. -- Brasília, DF : Instituto Walden4, 2026.

PDF

Other authors: Carlos Augusto de Medeiros, Cristiano Coelho, Diogo Conque Seco-Ferreira, Frederico Neves Condé, Lucas Ferraz Córdova

Translation of: Uma análise operante de discursos de ódio e desinformação (ISBN 978-85-65721-74-5).

Bibliography.

ISBN 978-85-65721-85-1

1. Aggressive behavior 2. Behavior - Analysis 3. Behavior - Psychological aspects 4. Communication - Political aspects 5. Communication - Social aspects 6. Disinformation 7. Fake news 8. Hate - Social aspects 9. Politics - Social aspects I. Moreira, Márcio Borges. II. Medeiros, Carlos Augusto de. III. Coelho, Cristiano. IV. Seco-Ferreira, Diogo Conque. V. Condé, Frederico Neves. VI. Córdova, Lucas Ferraz.

Indexes for systematic catalog:
1. Behavior : Analysis : Psychology 150

Catalog record prepared by the publisher

ISBN 978-85-65721-85-1
9 788565 721851

Instituto Walden4
Brasília, DF · 2026

The Idea for This Book

The idea for this book came from a social media post that Rafael Nunes Félix sent to Carlos Augusto de Medeiros. We record here Rafael's credit for the idea, with our thanks.

Rafael Nunes Félix is a psychologist who graduated from CEUB, with interests in research and clinical practice in behavior analysis. He taught at the Universidade de Brasília (UnB), where he gave classes in Personality Psychology from a behavior-analytic perspective, and he is currently a non-degree student in the Graduate Program in Behavioral Sciences at the same university. He studies delay discounting and behavioral resurgence, with a focus on basic behavioral processes and their applications.

About the Authors

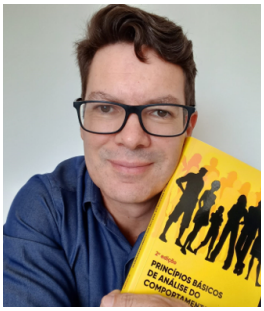
Márcio Moreira



Márcio Borges Moreira holds a doctorate in Behavioral Sciences from the Universidade de Brasília, and a master's degree in Psychology and a degree in Psychology from the Pontifícia Universidade Católica de Goiás. He teaches in the undergraduate program and in the Master's program in Psychology at the Centro Universitário de Brasília (UniCEUB), where he supervises clinical internships at the Centro de Formação de Psicólogos (CENFOR). He is also the director of the Instituto Walden4, a clinic specializing in ABA-based services for children and adolescents with autism spectrum disorder (ASD).

Coauthor of the book *Princípios Básicos de Análise do Comportamento* [Basic Principles of Behavior Analysis], he has published book chapters and scientific articles in Brazilian and international journals, with a focus on applied behavior analysis, verbal behavior, and evidence-based interventions. His current research projects investigate the automation of training for professionals and paraprofessionals who work with people diagnosed with autism, the use of technology and artificial intelligence in teaching psychological concepts, the formation and change of prejudiced attitudes, and the measurement of behavior in the delivery of psychological services.

Carlos Augusto de Medeiros



Dr. Carlos Augusto de Medeiros completed his doctorate with honors in the Behavioral Sciences program at the Universidade de Brasília (UnB) in 2003. He earned his master's degree in Psychology, also at UnB, in 1999, having completed his undergraduate degree in Psychology at the same institution in 1997. He served as an assistant and adjunct professor at the Universidade Federal de Minas Gerais (UFMG) from 2000 to 2003. At the Centro Universitário de Brasília (UniCEUB), from 2003 to 2024, he served as an undergraduate professor, supervisor of professional internships, and coordinator, professor, and advisor in the Master's program in Psychology. He has taught numerous modules in graduate courses in behavior analysis across Brazil,

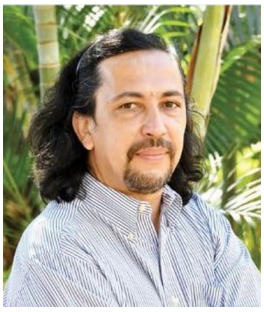
and he served as a clinical internship supervisor at the Instituto Brasiliense de Análise do Comportamento (IBAC). Over his extensive career, he has taught courses in behavior analysis, introduction to psychology, and behavior-analytic clinical practice, advised research projects, and supervised research and professional internships.

Since 1997, he has worked as a clinical psychologist, with an emphasis on individual therapy for adults and couples therapy. He is currently one of the directors of Pragmática: Psicoterapia e Cursos, a psychology clinic that offers regular continuing-education courses in clinical psychology.

He developed a model of behavior-analytic therapy called Pragmatic Behavioral Psychotherapy, a model grounded in the philosophy of radical behaviorism and committed to developing clients' independence and autonomy.

Throughout his long career, Dr. Carlos Augusto de Medeiros has devoted himself to training behavior analysts in Brazil. He is coauthor of the best seller "Princípios Básicos de Análise do Comportamento" [Basic Principles of Behavior Analysis]. He has published 22 articles, two books, and 43 book chapters. His publications deal with verbal behavior, equivalence relations, rule-governed behavior, behavioral clinical practice, and behavior analysis proper.

Cristiano Coelho



He holds an undergraduate degree, a master's degree, and a doctorate in Psychology from the Universidade de Brasília. He was a Substitute Professor at UnB (1999 to 2000), a Collaborating Professor at the Instituto de Educação Superior de Brasília (IESB) (2009 to 2010), State Survey Supervisor for IBGE, Brazil's national statistics institute, in Goiás (2010 to 2014), a Collaborating Professor in the MBA in Behavioral Economics at ESPM, and a member of the 2019–2020 board of the Brazilian Association of Behavioral Psychology and Medicine (ABPMC). He is currently an Adjunct Professor I at the Pontifícia Universidade Católica de Goiás, Coordinator of the Graduate Program in Psychology at PUC-Goiás (master's and doctoral levels), and a behavior therapist. He has conducted research and advised undergraduate research, master's, and doctoral students mainly on the following topics: choice behavior, self-control and impulsivity, social skills, and consumer behavior analysis.

Diogo Conque Seco-Ferreira



Diogo Conque Seco-Ferreira earned his degree in psychology from the Universidade de Brasília (UnB), where he also received a master's degree in Psychology and a doctorate in Behavioral Sciences, with research on consumer behavior, and he completed a postdoctoral fellowship in behavioral economics at Tilburg University, in the Netherlands. He is an associate professor in the Department of Psychology at the Universidade Federal de Sergipe (UFS).

He works at the interface between behavior analysis, economic psychology, and behavioral economics, with research on consumer behavior, the behavioral analysis of social phenomena, and the behavioral sciences applied to public policy. He is a coauthor of "Economia comportamental: uma introdução para analistas do comportamento" [Behavioral economics: An introduction for behavior analysts], published in the Revista Interamericana de Psicología.

Frederico Neves Condé



He holds a doctorate, a master's degree, and an undergraduate degree in Social, Work, and Organizational Psychology from the Universidade de Brasília (UnB). He is a researcher-technologist in Educational Information and Assessment at Inep (Brazil's National Institute for Educational Studies and Research), where he headed the Service for the Calculation of Indicators from Contextual Instruments (2022–2023), served as Director of Basic Education Assessment (2014–2015), and was one of the developers of the Saeb socioeconomic level index, Inse (2019 and 2021). He was a technical advisor to the Executive Presidency and the Ombudsman's Office of Anatel (Brazil's telecommunications regulator), a social information analyst at the MDS (Ministry of Social Development), a consultant to UNDP and UNESCO, and a council member of CRP-01 (the Regional Psychology Council for the Federal District). He has taught undergraduate and graduate courses (IESB and UniCEUB) and served as chief executive of Psicométrica Consultoria e Pesquisa. His areas of work are psychometrics and psychological, educational, and social assessment, with a focus on Item Response Theory, instrument development, and indicator modeling.

CV: http://lattes.cnpq.br/8338146219871041

Lucas Ferraz Córdova



Lucas Ferraz Córdova earned his degree in psychology from the Universidade Federal de Minas Gerais (UFMG) and holds a master's degree in Psychology and a doctorate in Behavioral Sciences from the Universidade de Brasília (UnB), with a dissertation on the effect of successive training on transfer between the mand and tact verbal operants. He is a professor at the Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM).

He works in behavior analysis, with an emphasis on verbal behavior, the philosophy of radical behaviorism, and the relations among language, context, and scientific explanation. He coordinates projects on the environmental and historical determinants of the behavior of explaining, which investigate the variables that affect the establishment and maintenance of the verbal repertoire, as well as experimental studies on verbal operant training, reinforcement history, and the effects of audiences on explanatory discourse.

Contents

Preface 13

Scope and limits 13
The reading rules 13
How to read the boxes and the terms 14
The map of the book 15
Reading order 16
References *16*

A Note on the Production Process 17

What came from the authors 17
What the AI did 18
Sample prompts 20
What we did afterward 25
Limits of the process 26
A behavioral reading of the process itself 26
Responsibility 27
The English edition 27

PART I **The Phenomenon and the Data** 28

Chapter 1 — Disinformation and Hate Speech: Definitions, Measures, and Data 29

1.1 Definitions, and what each one requires measuring 29
1.2 Prevalence: how much circulates and who shares 31
1.3 Political elites as a source 33
1.4 Political position and propensity toward false information: what the correlational studies show 34
1.5 Limitations: what the designs do not allow 37
1.6 What is known about Brazil 38
1.7 Hate speech: definitions, measures, and data 40
1.8 Rival interpretations and limits 41
Summary *42*
References *43*

Chapter 2 — What Fact-Checking Agencies Say (and What They Cannot Say) 46

2.1 Two series of contingencies 47
2.2 What each source counts 49
2.3 Why the proportion among checked statements is not the proportion among statements made 50
2.4 What Lupa allows one to say 52
2.5 What Aos Fatos allows one to say 54
2.6 Remarks and mentions: whose falsehood is it? 56
2.7 Do the agencies agree with one another? 56
2.8 What fact-checking cannot reach 57
2.9 Protocol for a comparable sample 58

Summary *59*
References *60*

PART II **Operant Analysis of the Discourses** 62

Chapter 3 — False Statements as Verbal Operants **63**

3.1 The politician's speech as verbal behavior 64
3.2 The analysis form 66
3.3 Tact or intraverbal? The continuum and the tact form 68
3.4 The distorted tact: when the listener's consequence displaces control by the event 71
3.5 Repetition: the strength of the repeated claim 74
3.6 Autoclitics of source and of certainty: what is added and what is removed 77
3.7 The statement as a rule for the audience 80
3.8 Cases: the same four statements, by the same method 81
3.9 Rival interpretations and how to decide among them 85
Summary *86*
References *87*

Chapter 4 — Political Attitudes and Prejudice as Equivalence Classes **90**

4.1 Attitude as evaluative verbal behavior 91
4.2 How a word acquires a function without contact with the referent 93
4.3 Transfer of function 97
4.4 Prejudice in the experiments 102
4.5 Attitudes toward Brazilian presidents 106
4.6 "Us" and "them": relations of opposition and comparison 120
4.7 Attitude change: reorganization of classes and comics 122
4.8 Measures, rival interpretations, and limits 131
Summary *134*
References *135*

Chapter 5 — Hate Speech: Disguised Mands and Aversive Control **140**

5.1 What the speaker's behavior produces: approval, removal of threats, and functions that vary with context 141
5.2 The disguised mand: the opinion that functions as an instruction 143
5.3 Aversive control directed at a group 146
5.4 The group label and the transfer of aversive function 148
5.5 From speech to act: what the data allow us to state 149
5.6 A case with external classification: the 2011 remark and what came after 154
5.7 Rival interpretations and limits 156
Summary *157*
References *158*

Chapter 6 — Digital Bubbles as Verbal Communities and the Avoidance of Countercontrol **161**

6.1 The bubble as a verbal community: audience and source of reinforcement 162
6.2 What the data say about echo chambers (and where they diverge) 164
6.3 "Virtual realities": when verbal control replaces contact with the event 168
6.4 Countercontrol and avoidance of countercontrol 169
6.5 The Brazilian case: WhatsApp, elections, and the pandemic 174

6.6 Who likes, who replies:
an exploratory analysis of two political videos on Instagram 177
6.7 Semantic analyses on social media 192
6.8 Rival interpretations and limits 193
Summary *195*
References *196*

Chapter 7 — Voting as Verbal Behavior 199

7.1 Voting under the definition of verbal behavior 200
7.2 Mand, tact, or intraverbal? Possible functions of the vote 202
7.3 The secret ballot as removal of the audience 203
7.4 Declaring and voting: verbal attitudes and acts 205
7.5 Social consequences of voting: the field experiments 210
7.6 Candidates' names and numbers in equivalence classes 213
7.7 Rival interpretations and limits 215
Summary *216*
References *217*

PART III The Ecology of Information 219

Chapter 8 — Attention Economy: The Algorithm as a Programmer of Contingencies 220

8.1 Attention as a reinforcer 221
8.2 What the data show about emotional, moralized, and negative content 223
8.3 The algorithm as programmer: how large an audience each emission receives 226
8.4 Intermittent reinforcement and engagement 230
8.5 Experiments on the algorithm itself: what changed and what did not 235
8.6 Rival interpretations and limits 238
Summary *239*
References *240*

Chapter 9 — The Producer-Audience Cycle 243

9.1 Opinions can be reinforced: the experiments of 1956 to 1969 244
9.2 What the producer receives: likes, shares, and reach 250
9.3 What the audience does and what maintains it 253
9.4 The cycle: each side shapes the other 255
9.5 Moral outrage learned through reinforcement: the evidence 256
9.6 The politician as content producer 259
9.7 The metacontingency of the ecosystem:
producer, audience, platform, and advertiser 260
9.8 Rival interpretations and limits 262
Summary *263*
References *264*

Chapter 10 — The Overton Window: What Can Be Said Without Being Punished 267

10.1 The concept, its origin, and the proposed translation 268
10.2 From the policy window to the statement window: what shifts the edge 271
10.3 Information about what others tolerate: the Pittsburgh experiment 272
10.4 Two elections, two measures: Brazil, 2018, and the United States, 2016 275

10.5 Politicians and social media: who speaks from above and what they receive 279
10.6 Sanction, countercontrol, and what holds the edge in place 281
10.7 How to situate a statement in relation to the window 284
10.8 Rival interpretations and limits 285
Summary *287*
References *288*

Chapter 11 — Infodemiology: How Information Spreads and What Is Measured 290

11.1 Infodemiology and infoveillance 291
11.2 Measures of diffusion: reach, exposure, reaction, and cascade structure 293
11.3 Does falsehood spread faster than truth? 296
11.4 Epidemiological models applied to information 299
11.5 The COVID-19 infodemic, worldwide and in Brazil 302
11.6 What infodemiology proposes to do 305
11.7 Questions for reading a diffusion study 307
Summary *307*
References *308*

PART IV Interventions 310

Chapter 12 — Interventions: What Can Be Changed, at Four Levels, and with What Evidence 311

12.1 Correcting afterward: correction, warnings, and the "backfire effect" 313
12.2 Correcting beforehand: inoculation as discrimination training 316
12.3 Stimuli that signal accuracy 320
12.4 In the face of the speaker: counterspeech, public commitment, and source autoclitics 323
12.5 Reducing prejudice through equivalence and comics 326
12.6 Changing platform contingencies 330
12.7 Cultural design and the ethical limits of those who design 334
Summary *337*
References *338*

Appendix A — Discourse Analysis Form 341

Appendix B — Concepts in Use 343

Appendix C — References 347

Appendix D — Source Verification 357

D.1 How a claim comes to be "verified" 357
D.2 Numbers 358
D.3 Versions read that are not the journal version 358
D.4 Sources not read 359
D.5 Methodological decisions that affect the reading 365
D.6 Recorded procedural slips 366

Preface

This book brings together a question and a method. The question is what maintains, in terms of contingencies, what politicians, platforms, and audiences say and pass on when the statement is false or is classified by someone as hate speech. The method is that of behavior analysis applied to verbal behavior: instead of inferring what the speaker had in mind, one asks which variables control the emission and what it produces in the listener. Skinner (1957) provides the vocabulary; the experimental literature on attitudes, equivalence, and cultural contingencies provides what is known and what is not yet known.

It was written for professionals who already know behavior analysis: behavior analysts, psychologists, researchers. That is why there is no part on fundamentals. Each concept enters, in a "Concept in use" box, at the point where the analysis comes to need it, and Appendix B gathers them together with the chapter in which each one was introduced.

Scope and limits

No utterances are classified as false here, no one is singled out as a liar, and no political camp is defended. The one who classifies is a fact-checking agency, a court decision, a platform, or a study, and the text says which. Nor was any hypothesis tested: studies by others were read, and wherever an operant reading is proposed, the text says that it is a reading. What can be taken from here is a way of reading the data and of formulating the question that is missing, not an answer about who is more dishonest.

Two consequences appear in every chapter. The first is that the classification of a verbal operant (mand, tact, intraverbal, distorted tact) is always interpretation, never data: the text says which controlling variables support it and what the rival interpretation would be. The second is that coverage is uneven. A large share of the studies are in English and from the United States, and they come mostly from Twitter and from Covid-19; for Brazil there are few. Each chapter says what this does and does not allow one to conclude.

The reading rules

Five rules hold from beginning to end.

The first is the same method for all political camps. Wherever what a politician said is analyzed, the mirror case from the other camp is sought with the same criterion, and when it is not found, the text says there is none. The comparable-sample rule of Chapter 2 (Section 2.9) was set before any data collection, so that the choice of cases would not depend on the result; the collection was not carried out, and the chapter says why.

The second is that falsity always belongs to the agency that classified it. “False” appears with the name of the agency, the label it used, and the date, and never as a property that the writer attributes to the sentence.

The third is that “hate speech” is an external label. Only a statement classified by an external body, with that body named, is called that. Without such a classification, the function of the statement is described, and the statement is not labeled.

The fourth is the separation between data, interpretation, and hypothesis. Every analysis of data or of discourse comes with a box in three parts: what the source shows, what is read into it, and what remains a hypothesis, together with what it would take to test it.

The fifth concerns sources. Only what was read in full text enters as a source, and each claim attributed to a source was located in it and read in context; the spreadsheets for this work are described in Appendix D. Where a source could not be obtained by legal means, the text states that it was not read, and it is not cited as verified. Page numbers cited are those of the journal when the text that was read includes them; when it does not, the citation goes without a page number, and the chapter notes which version was read (accepted manuscript, working paper, XML text).

How to read the boxes and the terms

There are three types of box. The “Data · Interpretation · Hypothesis” box closes the analyses. The “Concept in use” box presents a concept from the literature, with its source and the point at which it begins to be used; when a chapter returns to it, the box says “revisited” and refers to the chapter of origin. The “Proposed definition” box marks a term that is proposed here and that no source read uses, such as an audience’s window (Chapter 10) and the distinction between reach, exposure, and reaction (Chapter 11); “avoidance of countercontrol” (Chapter 6) is a proposal made here, building on Delprato (2002), and always appears as a hypothesis.

Citations follow the APA standard: "Author (year, p. X)." The references for each chapter are consolidated in Appendix C. In the text, "was read" and "was not read" mean exactly this: the full text was or was not read by those who wrote this book.

The studies cited report statistics, and the text explains them the first time it uses them; this paragraph is for keeping them at hand. The *p* value is the probability of obtaining a difference as large as the one observed, or larger, if there were in fact no difference; by convention, a *p* below .05 leads the difference to be called "significant," which indicates that it is unlikely to be due to chance, not that it is large or important. *t* and *F* are the statistics of the tests that produce the *p*. Cohen's *d* and *h* measure the size of an effect, the difference between means in the first case and between proportions in the second, and values near 0.2 are small. The *odds ratio* (OR) compares the odds of an outcome in two groups: 1 means equal odds. The standard deviation is the typical variation of a measure, and the 95% confidence interval is the range of plausible values for the estimate. R^2 is the fraction of the variation in the data that a model explains, and accuracy, precision, recall, and *F*1 describe how often a text classifier gets it right. When a study has no control group, or measures only what people report, the text says so.

The map of the book

The book is divided into four parts and four appendices. Part I, "The Phenomenon and the Data," has two chapters. Chapter 1 brings together the definitions and the prevalence and correlational studies, with the limitations of each measure, and establishes the rule that what is measured is a function of the instrument. Chapter 2 deals with fact-checking agencies: what the proportion of false statements published by an agency does and does not measure, and why it cannot be compared with another agency's proportion without the caveats set out in the chapter.

Part II, "Operant Analysis of the Discourses," has five. Chapter 3 is the core: it presents the analysis form and the verbal operants that the following chapters use, and applies the form to four fact-checked utterances. Chapter 4 deals with attitudes and prejudice as equivalence classes. Chapter 5 deals with hate speech as a disguised mand and as aversive control; Chapter 6, with digital bubbles as verbal communities and with the avoidance of countercontrol; Chapter 7, with voting as verbal behavior.

Part III, "The Ecology of Information," has four. Chapter 8 deals with the attention economy and the algorithm as a programmer of contingencies; Chapter 9, with the producer–audience cycle; Chapter 10, with the Overton window,

defined by what an audience punishes; Chapter 11, with infodemiology and measures of diffusion. Part IV has a single chapter, Chapter 12, on the interventions that follow from the analysis, at four levels, and the evidence for each. The appendices contain the analysis form (A), the concepts in use (B), the consolidated references (C), and source verification, with its gaps (D).

Reading order

Chapters 1 and 2 can be read on their own by those who want only the data and the critical reading of fact-checks. Chapter 3 is the foundation for the others: anyone who will read the analysis of utterances and of hate speech, in Chapters 5 and 7, should go through it and through Chapter 4. Apart from that, the chapters are self-contained, with explicit cross-references to whatever they draw from another chapter. Chapter 12 is best read last, because it uses the discrimination-training framework, the distinctions of Chapters 8 and 9, and the definition of reach, exposure, and reaction from Chapter 11.

Those who use the book as an analytical tool will find ready-made questions at four points: the form in Chapter 3 (Section 3.2), reproduced in Appendix A, for reading an utterance; the questions in Chapter 10 (Section 10.7), for locating an utterance within a window; those in Chapter 11 (Section 11.7), for reading a measure of diffusion; and those in Chapter 12 (Section 12.7), for reading an intervention. The form is the one most used: it turns what the chapters describe into a sequence of questions about the utterance, the audience, and the consequences, and it always ends with the question that matters, which is what maintains the utterance.

References

Delprato, D. J. (2002). Countercontrol in behavior analysis. *The Behavior Analyst*, *25*(2), 191–200. https://doi.org/10.1007/bf03392057

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

A Note on the Production Process

This book was written with intensive use of generative artificial intelligence, and this note describes how. We, the six signatories (Márcio Moreira, Carlos Augusto de Medeiros, Cristiano Coelho, Diogo Conque Seco-Ferreira, Frederico Neves Condé, and Lucas Ferraz Córdova), are the authors: we defined the problem, the interpretive framework, and the criteria of analysis, and we answer for every published claim. The AI drafted the initial versions of all the chapters, of the Preface, and of Appendices A and B, and it carried out the verification of sources; we reviewed, altered, adjusted, and checked them between October 19 and 26, 2026. The direction of the AI, from the original request to the last revision request, was carried out entirely by Márcio Moreira.

We prefer to describe the process in detail rather than sum it up in a line of acknowledgment. Readers have the right to know how the text was produced in order to evaluate it, and a vague description would serve poorly both transparency and the work we actually did. For the same reason, we do not call this an "AI-generated" book: the expression poorly describes a process in which the problem, the method, and the final judgment were human from beginning to end.

What came from the authors

What guides the book predates the AI-assisted drafting and came from us, in the original request of October 10, 2026, and in the requests that followed.

- **The problem.** The choice of hate speech and disinformation as the object, with a focus on the statements of Brazilian politicians, and the central question: what a Skinnerian analysis can say about this phenomenon on the basis of the data that cyberpsychology and its neighboring fields (political science, communication, infodemiology) have already produced.
- **The interpretive framework.** An explicit Skinnerian position, grounded in verbal behavior (Skinner's *Verbal Behavior* and Ferster and Skinner's *Schedules of Reinforcement*, which we admitted as an exception to the rule that only articles are cited), combined with Bernard Guerin's contextual analysis. From the Guerin article we designated as the basis come attitudes and beliefs as verbal behavior maintained by verbal communities, the tact, intraverbal, and mand functions of one and the same statement,

and the role of autoclitics in the social negotiation of what is said; from another of his works come the analysis of knowledge as a social construction and the condition under which a small, cohesive group sustains what it says. Added to this are stimulus equivalence and transfer of function, developed in our own books on attitudes and on comics. This framework reached the tools through the original request, the base article, those books, and the copies of Skinner that we supplied.

- **The selection of the literature.** We set the rules: only scientific articles, plus our books on attitudes and comics and the two works by Skinner; only texts obtained by legal means, without Sci-Hub or equivalents; only what was read in full text is cited; sites that prohibit AI crawlers, such as those of Lupa and Aos Fatos (Brazilian fact-checking agencies), are not scraped in bulk. The list of candidates came partly from us (27 articles from our collection and 12 of our books) and partly from the AI (97 articles anticipated from the table of contents and 34 added after 20 targeted searches in OpenAlex). We approved the survey on October 10, 2026, and it reached 168 sources, of which 108 were read in full text and 60 remained pending (Appendix D.2 and D.4).
- **The criteria of analysis.** What counts as a contingency, as a social consequence, and as verbal control. Classifying a verbal operant (mand, tact, intraverbal, distorted tact) is always interpretation, never data. The same method applies to politicians of all camps; "false" is always the classification of a named agency; "hate speech" is the label of an external body; data, interpretation, and hypothesis are kept separate. And what we do not intend to assert: we do not offer robust data on the phenomenon, but readings that may guide future experiments in behavior analysis.
- **The structure.** The ten topics of the original request (data and correlational studies, fact-checking agencies, verbal operants, stimulus equivalence, digital bubbles, voting, the attention economy, the producer–audience cycle, infodemiology, and interventions), the chapter on the Overton window, which we added on October 11, and the approval of the four-part table of contents, on October 10.

What the AI did

We used Anthropic's Claude Sonnet 5.5 in Claude Code, an environment in which the model reads and writes the project's files, runs scripts, and queries open bibliographic databases, between October 10 and 26, 2026. The version history also shows Claude Opus 5.5 in tasks of repository organization and lay-

out, without drafting any of the book's text. The tools always worked from the interpretive framework and the instructions we supplied, and the person who directed them, from beginning to end, was Márcio Moreira.

- **Surveying and obtaining the sources.** It proposed candidates; checked authors, year, journal, volume, and pages in Crossref; obtained the texts by legal means (open access, repositories, versions posted by the authors, and the PDFs we attached); and extracted the text, with optical character recognition when the text layer was poor. What it did not obtain went onto a list of pending items and is not cited as verified.
- **Structure.** It proposed the detailed table of contents and the order of writing: Chapters 4, 9, 6, 8, and 10 first, because their sources were complete, and the others afterward.
- **Initial drafting.** Working from the framework, it developed the first versions of all the theoretical sections, of the Preface, and of Appendices A and B.
- **Application of the framework to the studies.** It produced the first Skinnerian readings of the selected studies: description of the procedures, identification of possible contingencies, rival interpretations and limits, and interpretation of the results, always with data, interpretation, and hypothesis kept separate.
- **Verification of claims.** It recorded each claim attributed to a source in a spreadsheet, with the verbatim sentences that support it; a script searched for those sentences in the full text, and the AI read the context and recorded the result and any caveats. A second script checked that every citation in the text has a reference and that every flagged claim exists and is verified. The spreadsheet has 1,207 entries (Appendix D.1). In the initial survey, the checking corrected authors, volume years, and pages that the AI had recorded from memory, and it led to the removal of a claim attributed to an article that does not make it.
- **Figures, references, and files.** It generated the figures by script and cropped, from the Instituto Walden4 books, the images of the studies that Chapter 4 describes; it assembled the references in APA style and the delivery files (.docx, HTML, and PDF).
- **Rewrites on request.** It rewrote the study descriptions in prose and redid Chapter 4 with the images from our studies and from the comics, in accordance with the requests described below.

The AI is not an author. By delegation from us, it decided many drafting questions (the order of the chapters, the numbering, the titles of chapters, the handling of gaps, what was left out when a source could not be obtained), and

each decision is recorded with its date and with the way to reverse it. Three terms that appear as the book's own definitions originated in the AI-assisted drafting: "avoidance of countercontrol" (Chapter 6), "an audience's window" (Chapter 10), and the distinction between reach, exposure, and reaction (Chapter 11). Each one is declared as a proposal in the text, and the evaluation of them is ours. The final word on each interpretation belongs to the authors.

The drafting took place in conversation sessions with the AI over the project's files, one chapter at a time, beginning on October 10, 2026. It started with the original request, which the repository keeps unaltered. The AI delivered in stages (the base article, the table of contents, the bibliographic survey, and only then the chapters), and we approved each stage before the next. At other moments we replied "you can decide for me" and "Continue writing the book and finish it. You can make all the decisions": the AI decided and recorded each decision, with its date, in a dedicated file, and the delegation did not count as approval of a chapter. We also asked for course corrections after reading the draft: that the text stop treating "the book" as the one who reads, says, and proposes; that section openings stand on their own without the heading; that the research descriptions, which were truncated and not very didactic, be revised paragraph by paragraph; and that Chapter 4 be redone with colorful, didactic figures and with the images from our studies, including those from the comics. Some of these requests appear verbatim in the next section. The complete record, with the original request, the dated decisions, the verification spreadsheet, and the version history, is in the project repository and can be consulted upon request.

Sample prompts

Almost all the requests were short messages, in everyday language, sent in the middle of the work; the exception is the original request, long and structured, which we reproduce in full. All of them were written by Márcio Moreira. The prompts were written in Portuguese and are given here in translation. The other examples are verbatim and come from different points in the process; in them the spelling has been corrected, and words that were added or changed are in square brackets. Each example is followed by what the AI did with it, according to the project's record.

The original request (October 10, 2026), in full. The headings of the parts of the request appear in bold.

> **ROLE**

You are a behavior analyst (Skinnerian framework, Experimental and Applied Behavior Analysis) with experience in verbal behavior, stimulus equivalence, and the analysis of cultural practices, writing for an audience of advanced undergraduate and graduate students in Psychology.

OBJECTIVE

Write the book "An Operant Analysis of Hate Speech and Disinformation," in Brazilian Portuguese, in language that is didactic and rigorously technical (behavior-analytic terminology, operational definitions, no explanatory mentalism). The focus is on false discourses and statements by Brazilian politicians.

WORKFLOW (carry it out in stages and stop for my approval at the end of each one)

1. First read the base article: https://pmc.ncbi.nlm.nih.gov/articles/PMC2733686/ and summarize how it will ground the analysis of verbal behavior.
2. Propose the detailed table of contents (chapters, sections, planned figures, planned sources).
3. Assemble the bibliographic survey (see "Source rules") and the verification table, before writing any chapter.
4. Write one chapter at a time, with figures, and deliver it as a document.
5. At the end, cross-check all citations and references.

CONTENT (add relevant topics that I have not listed)

a) Data and statistics on disinformation ("fake news") and hate speech, including correlational studies between susceptibility to disinformation and political orientation. Present the limitations (correlation ≠ causation, samples, measures, cultural context).

b) Data from fact-checking agencies (Aos Fatos, Lupa, etc.) on false or misleading statements by Jair Bolsonaro during his term and by Lula. Use the same criterion for both: same agency, comparable period, same classification methodology, explicit denominator (no. of statements checked), and a caveat about selection bias in fact-checking.

c) Analysis of the discourses in terms of verbal operants (mand, tact, echoic, intraverbal, textual, etc., and distorted tact/disguised mand), based on the base article.

d) Analysis in terms of stimulus equivalence: class formation, transfer of function, conditional discriminations, properties (reflexivity, symmetry, transitivity, equivalence), relation to the formation of prejudice and attitudes. Reference Márcio Moreira's books on attitudes and on comics and attitudes (walden4.com.br/livros).

e) Digital bubbles as verbal communities; creation of "virtual realities" through verbal control; avoidance of countercontrol.

f) Voting behavior as verbal behavior.

g) Attention economy; research on users' tendency to consume emotional/moralized content; algorithms as schedules of reinforcement.

h) Reinforcement cycle between content producer and audience (likes, shares, audience as generalized reinforcers), analysis of the contingencies on both sides.

i) Information epidemiology (infodemiology): diffusion, propagation rates, interventions.

j) Practical implications: cultural and educational interventions based on behavior analysis.

FIGURES

Illustrate every procedure or concept that would benefit from a diagram: conditional discriminations (MTS), equivalence classes and their properties, the three-term contingency, the producer–audience cycle, bubbles as verbal communities. Generate the figures as files (Graphviz/matplotlib/SVG), with numbered captions, vector-based, legible in black-and-white print, and original (do not reproduce third-party figures).

SOURCE RULES

- Cite only scientific articles, plus his books on attitudes and comics (explicit exception). Do not cite other books.
- Obtain the articles only by legal means: open access (PMC, SciELO, Unpaywall, OpenAlex, repositories), institutional access that I make available, or PDFs that I attach. Do not use Sci-Hub or equivalents. Whatever is not obtained legally goes onto a "PDF pending" list and is NOT cited as verified.
- For each article cited, read the full text and confirm that it contains the information attributed to it. Record in a table: reference | DOI | excerpt/page supporting the claim | status (verified / pending).
- Validate DOI, authors, year, journal, volume, and pages (e.g., Crossref). Never invent or complete a reference from memory; if you cannot verify it, mark it as pending.
- Make two verification passes (in-text citations ↔ reference list ↔ verification table). Use APA 7 format.
- Literal quotations: short, in quotation marks, with page number, and accompanied by an image (crop) of the passage from the original PDF. Prefer faithful paraphrase.

NEWS

Use news stories only from G1, CNN Brasil, BandNews, and Canal Meio. For each one: a short excerpt quoted with credit, URL, publication date and access date, plus a screenshot of the excerpt. Also record the date of each utterance analyzed and the corresponding fact-check.

TONE AND POLITICAL RIGOR

> Describe behaviors and contingencies, not "essences" of people or parties. Analyze politicians from different camps with the same method. Always separate verified data, analytical interpretation, and hypothesis. Avoid defamatory statements; attribute each classification of falsity to the agency that made it.
>
> **DELIVERY**
>
> Each chapter as a document (.docx), figures in a separate folder, the reference verification table in a spreadsheet, and a list of pending items (PDFs not obtained, data not confirmed).

What the AI did. From this request came the work in stages (base article, table of contents, bibliographic survey, chapters, and cross-check), the verification spreadsheet, the rule of citing only what was read in full text, and the separation of data, interpretation, and hypothesis in every chapter. The instruction calling for figures legible in black and white was later revised at our request: the figures in Chapter 4 became color figures on October 16, and the others on October 26, 2026.

Approving and moving on (October 10, 2026). Between stages, we replied with phrases such as these:

> Go ahead.
>
> Approved. Continue.

What the AI did. It treated each reply as approval only of the stage the question covered; points that were not answered remained open in the decision log.

Delegating decisions (October 10 and 14, 2026).

> You can decide for me.
>
> Continue writing the book and finish it. You can make all the decisions.

What the AI did. It decided what remained open (order of writing, numbering, handling of sources that were not obtained, titles) and recorded each decision, with its date and with the way to reverse it. The delegation did not count as approval of a chapter.

Adding scope (October 11, 2026).

> Add a chapter with a behavioral analysis of the Overton window, relating it to politicians' discourses and social media.

What the AI did. It wrote the chapter, now Chapter 10, and proposed the definition of "an audience's window" that the chapter uses, declared as a proposal in the text.

Answering a question from the AI, with an ambiguous answer (October 11, 2026). The AI had asked whether the source exception opened for Skinner also applied to Ferster and Skinner's work on schedules of reinforcement. The answer was:

> Fester, ok. Go ahead.

What the AI did. It read "Fester" as Ferster and Skinner, recorded that reading, and noted that, if a different book was intended, we had only to say so.

Correcting the writing after reading the draft (October 16, 2026).

> "The book" come[s] in as an entity in the book itself (e.g., "The first study the book read is a game"). There are section-opening paragraphs that begin abruptly and confusingly or as if the section heading [were] already part of the paragraph. Identify and fix these issues.

What the AI did. It rewrote about 185 sentences in an impersonal voice, across all the chapters, and redid the section openings so that they can be understood without the heading.

> There are research descriptions that are truncated, not very didactic. Revise this writing issue paragraph by paragraph.

What the AI did. It rewrote in prose the study descriptions in the twelve chapters and in the Preface, one by one, with the question, the sample, the procedure, the measure, the result, and the caveat of each study.

Asking for detail and images (October 16, 2026).

> In the chapter on attitudes and equivalence, use these two works in detail, using their images: [addresses of the two works]. Use colorful, very didactic figures. Redo the chapter in this tone.
>
> Insert, wherever possible and appropriate, references to: [addresses of eleven works]. Use images from these books, make everything well illustrated, including images from the comics.

What the AI did. It redid Chapter 4 with color figures, cropped images from the studies and the comics of the Instituto Walden4, and included the works in the references, with the address of each PDF. One of the addresses appeared twice in the list, and the AI read it as the next volume, recording the assumption so that we could confirm it.

What we did afterward

Between October 19 and 26, 2026, the authors worked on the text produced in the previous stages: they reviewed, altered, adjusted, and checked the chapters, the Preface, and the appendices, in a shared manuscript.

1. **Conceptual review.** We assessed whether each argument was consistent with the Skinnerian framework and with contextual analysis, and we discarded or redid those that were not.
2. **Alteration and adjustment.** We corrected conceptual errors, inferences unsupported by the data, and extrapolations, and we rewrote the passages that called for a different formulation.
3. **Checking.** The checking of each claim against the full text of the source was done first by the AI, with the procedure described in Appendix D; we checked the texts during that same week of review. The 60 sources we could not obtain are not cited as verified, and where the subject would require the source, the chapter says that it was not read.
4. **Final editing.** We unified terminology, style, and the connections between chapters.

The box below summarizes the stages by chapter.

BOX

Stages of the work, by chapter

Chapter	First version by the AI	Works in the chapter's list	Verified entries	Review and checking by the authors
Preface	Oct. 15, 2026	2	0	Oct. 19–26, 2026
1. Definitions and data	Oct. 14–15, 2026	21	48	Oct. 19–26, 2026
2. Fact-checking agencies	Oct. 10, 2026	10	39	Oct. 19–26, 2026
3. False statements	Oct. 14–15, 2026	13	92	Oct. 19–26, 2026
4. Attitudes and equivalence	Oct. 10, 2026; redone Oct. 16, 2026	39	300	Oct. 19–26, 2026
5. Hate speech	Oct. 14–15, 2026	13	45	Oct. 19–26, 2026
6. Digital bubbles	Oct. 11, 2026	13	87	Oct. 19–26, 2026
7. Voting	Oct. 14–15, 2026	13	54	Oct. 19–26, 2026

Chapter	First version by the AI	Works in the chapter's list	Verified entries	Review and checking by the authors
8. Attention economy	Oct. 11, 2026	12	80	Oct. 19–26, 2026
9. Producer–audience cycle	Oct. 10, 2026	13	123	Oct. 19–26, 2026
10. Overton window	Oct. 11, 2026	11	80	Oct. 19–26, 2026
11. Infodemiology	Oct. 14–15, 2026	12	97	Oct. 19–26, 2026
12. Interventions	Oct. 14–15, 2026	27	162	Oct. 19–26, 2026

Note. The first-version dates are those of the first record of each file in the project history; Chapter 10 was begun on Oct. 11, 2026, under a different number and was renumbered on Oct. 14, 2026. "Verified entries" are the claims and data from sources recorded in the verification spreadsheet for the chapter, each with the passage read in context, before the authors' review and before Section 6.7. On Oct. 16, 2026, we requested two writing revisions of all the chapters (the text without "the book" as an agent, and the study descriptions in prose), and Chapter 4 was redone. Appendices C and D are generated by script from the spreadsheet.

Limits of the process

Two limits deserve to be stated. The first is that verification confirms that the passage attributed to a source is in that source's text, on the page indicated, not that the operant reading made of it is correct; that is the part the chapters call interpretation or hypothesis, and the part that depends most on our judgment. The second is that language models make errors when working from memory, and the AI itself recorded, at the start of the work, data that Crossref and the text of the sources contradicted; for that reason nothing entered the book without the checking described in Appendix D, and where a source could not be read, the text says that it was not read.

A behavioral reading of the process itself

For behavior analysts, the question of authorship has a well-known answer, and it does not change because of AI. In a 1972 essay, "On 'having' a poem," Skinner compares writing a poem to having a baby: the poet does not create out of nothing but is the locus where genetic and environmental variables come together. What we call creativity is variation and selection, and much of the writer's work lies in editing, that is, in selecting what remains.

We do not invoke Skinner to diminish the role of the AI or to inflate our own. The analogy serves to describe what actually changed: the AI greatly expanded the source of variation, and the authorial work became even more concentrated in the arrangement of conditions (the problem, the framework, the instructions) and in selection (what was accepted, corrected, rewritten, or discarded). This is a description of the process, not a justification of it; if readers disagree with it, the description given in the preceding sections still stands on its own.

Responsibility

The authors have reviewed the content of this book and take responsibility for it, including the parts whose first draft was produced with AI. Any remaining errors are ours.

The English edition

This edition was translated from the Brazilian Portuguese edition (ISBN 978-85-65721-74-5) with the same tools, and Márcio Borges Moreira signs as translator, with the agreement of the other five authors. Literal quotations from works written in English were restored to the authors' own words wherever the source could be consulted; quotations from Portuguese-language sources were translated. The figures were redrawn from the same data, with English labels. Another instance of artificial intelligence then compared the translation, sentence by sentence, with the original (adversarial review), and the errors found were corrected. Brazilian examples, institutions, and fact-checking agencies were kept as they are, with a short gloss the first time they appear.



PART I

The Phenomenon and the Data

1 Disinformation and Hate Speech: Definitions, Measures, and Data

2 What Fact-Checking Agencies Say (and What They Cannot Say)

CHAPTER

1

Disinformation and Hate Speech: Definitions, Measures, and Data

Allcott and Gentzkow's (2017) figures on the 2016 U.S. presidential election show, from the outset, the problem that organizes this chapter. In the database of fake news stories they assembled, 115 pro-Trump stories were shared 30 million times on Facebook, and 41 pro-Clinton stories, 7.6 million times. The same authors estimate that the average American adult saw and remembered 1.14 fake news stories, and they state explicitly that they do not assess whether those stories decided the election. A hasty reader folds these numbers into a single sentence, yet they answer different questions: the first measures what circulated; the second, what someone saw and remembered; neither measures what changed in what anyone did.

This chapter systematizes the accumulated knowledge on the amount and the forms of circulation of disinformation and hate speech, as well as on the profile of those who produce and share them, making explicit the metrics employed and their methodological limits. It is the least analytical chapter in the book: it does not propose an operant reading; rather, it shows the data with which the following chapters work and what each of them requires one to say only in a low voice. The rule of the whole book applies here: before reading the number, ask what the instrument controls.

1.1 Definitions, and what each one requires measuring

Every definition of "false information" carries a built-in measurement requirement. Lazer et al. (2018) define fake news as fabricated information that imitates journalistic content in its form, though not in its organizational process or its intent; it is a subcategory of misinformation in the broad sense, that

is, information about the state of the world that is wrong or misleading, and the outlets that carry it do not have the editorial norms and procedures of the news media. If what defines fake news is the publisher's intent and process, they write, prevalence should be measured publisher by publisher, not story by story. Allcott and Gentzkow (2017) define fake news as a news article that is intentionally and verifiably false and that can mislead the reader, and they exclude unintentional reporting mistakes from the definition. Both definitions require, in principle, knowing the intent; none of the measures that use them observes it.

Studies of sharing solve the problem by trading the item for the source. Guess et al. (2019) take fake news to be false or misleading content deliberately made to look like news, usually to earn advertising revenue, and they check the links each respondent shared against lists of fake news publishers compiled by journalists and scholars. Lasser et al. (2022) use the score that NewsGuard, a service that rates the trustworthiness of news websites, assigns to the domain (from 0 to 100, with below 60 "not trustworthy") as an estimator of the trustworthiness of the shared item. This is an explicit convention, that of inferring the item from the source, and not a hidden flaw; but it changes what the number says: the proportion of links to untrustworthy domains is not the proportion of false claims.

Experimental studies change the object once more. Rathje et al. (2023) write that misinformation can refer to fabricated news stories, false rumors, conspiracy theories, or disinformation, and what they measure is participants' judgment of true and false headlines. And Ricard and Medeiros (2020), in a note on the Brazilian case, distinguish false information that the person disseminating it believes to be true from false information that the person disseminating it knows to be false—two points on a "spectrum of 'information disorder'"—without presenting any empirical measurement of their own. The distinction is one of intent, and Table 1.1 shows that almost none of the measures observes it.

The terms are used as the sources use them. "False information" and "disinformation" (in the Portuguese original, desinformação) appear, with no claim to a definition of our own, for what the source calls misinformation, and intent enters only when the source measures it. As for "hate speech," the rule is the one followed in the rest of the book: it is an external label, applied only to a statement classified by an external body, with that body named; the studies in Section 1.7 use the expression as a category of their own measure, and here it is taken as they present it.

TABLE 1.1

Six definitions of false information and what each one requires measuring

Source	Term and definition	What the definition requires measuring	What the study measured
Lazer et al. (2018)	Fake news: fabricated information that mimics the form of journalism, not its process or intent	The publisher's process and intent	No measure of its own: an essay proposing measurement by publisher rather than by story
Allcott and Gentzkow (2017)	Fake news: an intentionally and verifiably false article that could mislead	Intent and veracity, story by story	Shares of 156 stories classified as false by fact-checkers in the three months before the election
Guess et al. (2019)	False or misleading content dressed up as news, often for ad revenue	Origin of the content	Links shared on Facebook, matched against lists of fake news publishers
Lasser et al. (2022)	Untrustworthy information, defined by the domain	Trustworthiness of the source, not of the item	NewsGuard score below 60 for the link's domain
Rathje et al. (2023)	Misinformation: fabricated news stories, rumors, conspiracy theories, or disinformation	Veracity of the headline	Participants' judgment of true and false headlines, with and without incentives
Ricard and Medeiros (2020)	Misinformation (the disseminator believes) and disinformation (the disseminator knows), citing another work	What the disseminator knows	No measure of its own: a commentary, with no empirical data

Note. Our own descriptions, based on the passages cited; the column "What the study measured" gives what each study operationalized, not what it declared it wanted to measure.

1.2 Prevalence: how much circulates and who shares

Prevalence can be investigated through three central questions, and each requires its own measuring instrument: how much circulates, who shares, and where the content comes from. The first question concerns how much circulates. For the 2016 U.S. presidential election, Allcott and Gentzkow (2017) assembled 156 news stories about the election that leading fact-checking sites had rated false during the three months preceding the vote. The 115 pro-Trump stories were shared 30 million times on Facebook, and the 41 pro-Clinton stories, 7.6 million times. Sharing, however, is not the same as having seen, and the authors tried to estimate what people saw: in an online survey of 1,200 people, conducted after the election, they asked whether respondents

remembered seeing each of 14 false headlines that had circulated. About 15% said they had, but about 14% also said they had seen placebo headlines—untrue headlines invented by the authors that never actually circulated. Because the placebo headlines could not have been seen, the authors take the difference between the two percentages as a measure of true recall and, projecting it onto the full set of stories in the database, estimate that an average adult had seen and remembered 1.14 fake news stories. In the same survey, only 14% of American adults regarded social media as the most important source of news about the election.

The second question is who shares, and the study by Guess et al. (2019) comes closest to an answer based on a record of behavior rather than on self-report alone. The authors administered an online survey to a representative sample of 3,500 Americans and asked each respondent to authorize access, through an application, to data from their own Facebook account, including their posts with external links; the analysis covers those who consented and whose data could be linked to the survey. The links shared in 2016 were then compared with the lists of fake news publishers that journalists and academics had compiled, which avoids the known biases of self-reports of online activity. The main result is one of rarity: about 8.5% of the respondents with linked data shared at least one story from a fake news domain, and over 90% shared none. It is not that people share little in general: during the period, 26.1% of respondents shared 10 to 100 links of any kind, and 61.3% shared 100 to 1,000. Among those who shared fake news, there were more Republicans: 38 versus 17 Democrats in absolute numbers, and, in the sample, 18.1% of the Republicans against 3.5% of the Democrats. The most robust finding, the authors say, is age: respondents aged 65 or older shared, on average, 0.75 stories from fake news domains, more than twice the second-oldest age group (0.26), and the difference persists when ideology and education are held constant.

Two caveats from the authors themselves accompany these numbers. The first is that the pattern by ideology is consistent with two facts: most fake news had a pro-Trump slant, and people tend to share what they approve of. The finding, therefore, may not mean that conservatives are more inclined to share fake news when they come across it. The second is that the authors have no measure of whether respondents knew the content was false (Guess et al., 2019). Lazer et al. (2018) add a limit of method. They cite a study that examined how a set of high-profile fake news stories spread and concluded that the average American came across one to three of them in the month before the election; the estimate is probably conservative, because the study tracked

only 156 stories. It should be stressed that counting the accounts that received or reposted a post does not reveal how many people read it, nor whether that stimulus had any effect on their actions.

In short, as Lazer et al. (2018) put it: exposure does not equal impact, and the literature contains essentially no assessment of how exposure to fake news affects political behavior over the medium and long term. Mosleh and Rand (2022) note a similar distinction on the side of those who share: exposure and sharing are, at bottom, distinct constructs; most people pass on only a very small share of what they see, and looking at what someone shares offers a very partial view of that person's information environment.

1.3 Political elites as a source

The third question about prevalence—where false information comes from—shifts the focus from the general public to political elites. Mosleh and Rand (2022) argue that, after the attention paid to news from obscure publishers, a different form of misinformation came to draw attention: campaigns coordinated by public figures and organizations, which they call elites. To study it, they built two scores. The first is a falsity score for each of 816 elites, calculated from PolitiFact's professional fact-checks of the veracity of their statements. The second is an exposure score for each Twitter user, calculated from the falsity scores of the elites that user follows. They then compared exposure with what the user shares. The quality of the news shared was measured by the trustworthiness of the domain, assessed in two ways: by the average rating of eight professional fact-checkers and by a sample of 970 Americans selected by quotas, with the ratings of supporters of the two parties balanced. The user's ideology was estimated from the ideology of the accounts the user follows. The greater the exposure to misinformation from elites, the lower the quality of the news the user shares, and the further to the right the user's estimated political leaning. This is a correlation between measures, and the authors themselves say that the approach relies on the immediate network of accounts followed and does not capture exposure beyond them, such as through retweets by connections or algorithmic recommendations.

Lasser et al. (2022) directly analyze the posts issued by legislators. The authors collected the tweets, excluding retweets, of members of the U.S. Congress, the German parliament, and the British parliament between 2016 and March 2022, extracted the links, and gave each link the NewsGuard score for its domain, from 0 (very poor) to 100 (exceptional quality journalism); below 60, the domain is considered "not trustworthy." Republicans shared 9.1 times more links to untrustworthy domains than Democrats: 3.86% of links, versus 0.43%.

In Germany, the CDU/CSU (Christian Democrats) shared 0.18% and the SPD (Social Democrats) 0.03%; in the United Kingdom, the Conservatives 0.25% and Labour 0.05%, and the authors note that these figures rest on very low counts. The Republicans' proportion of untrustworthy links rises from 2.4% in 2016–2018 to 5.5% in 2020–2022, while the Democrats' remains at 0.4%; the result holds with a second, independently compiled set of domain trustworthiness ratings that does not come from NewsGuard. The divergence between the parties appears to be peculiar to the United States: it does not appear in Germany or in the United Kingdom.

What the authors propose as an explanation is a debated possibility: partisan motivation, since derogatory content about the opposing political group increases sharing intentions among partisans. The authors also suggest that the pattern among legislators may contribute to the asymmetry observed among the public in two ways: by providing misinformation to partisans and by legitimizing the sharing of untrustworthy information. NewsGuard has scores for only part of the links, and the study reports the coverage: 46.5% of the links from the U.S. Congress, 58.8% from the German parliament, and 39.2% from the British one, after excluding social media websites and search engines. What has no score is not measured.

1.4 Political position and propensity toward false information: what the correlational studies show

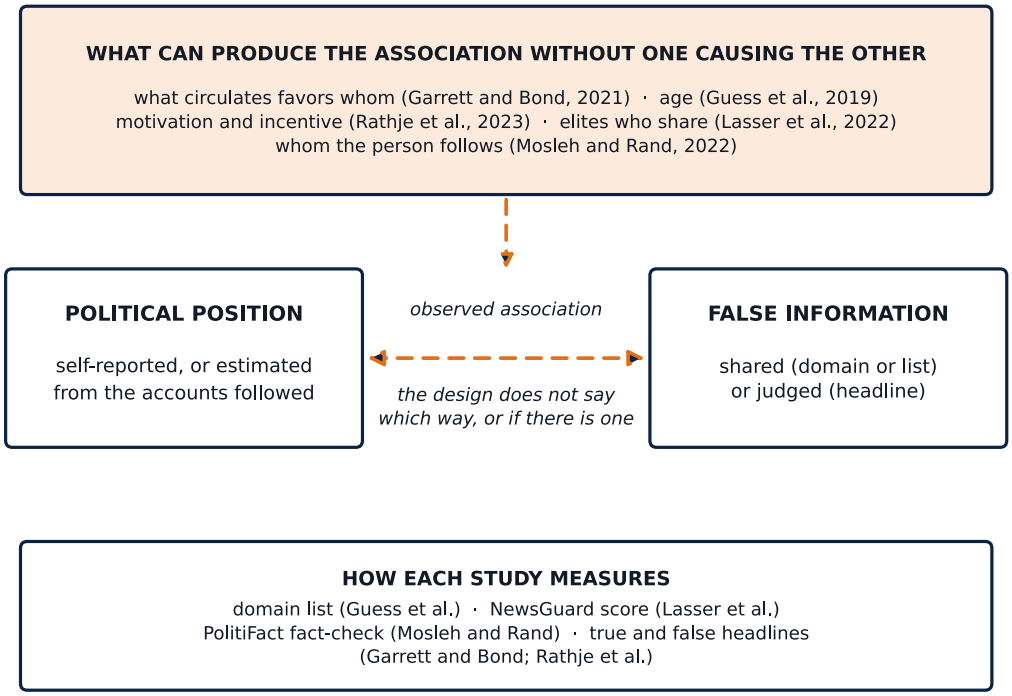
The recurring question in the literature is whether there is a partisan asymmetry: whether those who position themselves in a given camp share or believe false information more. The verified studies answer with associations, and Table 1.2 brings them together with what each design does not allow one to conclude. Figure 1.1 illustrates the limits of this inference.

Garrett and Bond (2021) took a methodological precaution that is unusual in the field. The evidence that conservatives hold more false beliefs rested on a relatively narrow set of topics, often chosen on purpose to mirror falsehoods pushed by conservative political elites. The authors instead left the choice of claims to an external criterion: every two weeks, a service identified the 20 political news stories that had spread most widely, half of them true and half false, and a 12-wave panel with a large representative sample of Americans rated, in the following week, short statements about those stories, over a period of six months. From the responses, they calculated two indices, following signal detection theory: sensitivity, which is the ability to distinguish true from false statements, and response bias, which is the propensity to label all statements as true or as false. In U.S. survey data, conservatives distinguished

truths from falsehoods less well than liberals did, and the authors attribute part of the difference to the information environment: prominent true claims tend to advance the issues and candidates that liberals favor, and false narratives tended to favor conservative causes, so that, even if they were equally biased, liberals would perform better in such an environment.

FIGURE 1.1

What an association between political position and false information shows, and what can produce it without one causing the other



Original diagram, no data. Dashed: a relation that correlational designs do not disentangle.

Note. Original diagram by the author, based on Garrett and Bond (2021), Guess et al. (2019), Rathje et al. (2023), Lasser et al. (2022), and Mosleh and Rand (2022). It does not represent data. Dashed line: a relation that correlational designs do not disentangle.

Rathje et al. (2023) add experiments in which the consequence of being right could be varied. In four experiments with 3,364 participants from the United States, people read true and false political headlines and answered whether the claim in the headline was accurate; a randomly assigned half were told they would receive a bonus of up to one U.S. dollar depending on their correct answers, and the other half answered the same questions without an incentive. The authors call partisan bias the difference between the number of headlines favorable to the participant's own party and the number of headlines unfavorable to it that the participant rated as true. Financial incentives for accuracy reduced partisan bias in headline judgments by about 30%. Conservatives were less accurate than liberals at distinguishing true from false head-

lines, but the incentives reduced the gap by 52%; and incentivizing participants to pick out the news their political allies would like reduced accuracy. Part of what appears as a difference in belief, therefore, seems to depend on what is reinforced: being right or pleasing.

TABLE 1.2

Five studies on political position and false information: what they measured and what they do not allow one to conclude

Study	Sample and country	Measure of false information	Measure of political position	Result	What it does not allow
Guess et al. (2019)	Respondents to a representative survey linked to Facebook data; U.S., 2016	Links from fake news domains, by list	Self-reported (party, ideology)	Rare (8.5%); more common among Republicans and among older people; age is the robust finding	Attributing the difference to greater propensity: it may reflect the pro-Trump supply
Garrett and Bond (2021)	U.S. survey, claims over 6 months	Judgment of true and false claims	Self-reported	Conservatives discriminated less well; part attributed to the information environment	Concluding that there is greater bias: the result depends on the set of claims
Mosleh and Rand (2022)	816 elites and Twitter users, U.S.	Falsity score from PolitiFact fact-checks	Estimated from whom the user follows	Exposure to elites with high falsity scores correlates with lower news quality and with conservative ideology	Causality; a single fact-checker; exposure beyond those followed
Lasser et al. (2022)	Legislators in the U.S., Germany, and the United Kingdom, 2016–2022	NewsGuard score of the link's domain	Party	Republicans share more, and more over time; Germany and the United Kingdom, low counts	Inferring the veracity of a statement from the domain of the source; coverage limited to 39% to 59% of links

Study	Sample and country	Measure of false information	Measure of political position	Result	What it does not allow
Rathje et al. (2023)	Four experiments, n = 3,364, U.S.	Judgment of true and false headlines	Self-reported	Incentives reduced partisan bias by about 30% and the gap by 52%	Generalizing beyond headlines and online participants

Note. Data compiled from the original articles cited. The restrictions in the last column make explicit what the methodological designs cannot isolate.

1.5 Limitations: what the designs do not allow

The studies gathered in the previous sections share three limits, and it is through them that the chapter's numbers should be read. Of the five studies in Table 1.2, four are about the United States and the fifth compares it with Germany and the United Kingdom; the data of three of them come from a single platform, Facebook or Twitter. The first limit, therefore, is one of sample and context: the findings come from the American two-party system, in specific elections, and Lasser et al. (2022) say that the divergence they observe appears to be peculiar to the United States. The second is one of cause: none of these designs separates what a person sees, what is offered to that person, and what the person chooses, and Figure 1.1 brings together the variables that can produce the association without political position and false information causing each other. The third is one of measurement, and it deserves a box, because it is what is borrowed from behavior analysis to read all the others.

CONCEPT IN USE

Verbal behavior, audience, and self-report

Verbal behavior is behavior reinforced through the mediation of other persons (Skinner, 1957); the audience is the discriminative stimulus in the presence of which that behavior is characteristically reinforced. Self-report—that is, responding to a scale, a survey question, or an interview—is verbal behavior, and it is under the control of who is asking and of how the question is asked. For Guerin (1994), attitude scales and survey questions depend on who is asking and on the verbal culture in which the survey takes place, because attitudes always depend on their social consequences. The classic example is that of Hildum and Brown (1956): in a telephone survey of 40 male Harvard students, the interviewer said "Good" after answers for, or against, the General Education program, and the "Good" shifted the answers in the reinforced direction; only 8 of 20 subjects noticed what the interviewer was doing, and all of them denied that their answers had been influenced by it.

The consequence for this chapter is direct. Table 1.2 contains three rows in which the data are a response to a survey or to a judgment task (Garrett and Bond, Rathje et al., and, in part, Guess et al., in which ideology is self-reported), and a response to a survey is behavior under the control of the situation in which it was given. The result of Rathje et al. illustrates precisely this dynamic: when the consequence of being right changes, the verbal judgment changes. A second measurement problem is that of stimulus selection. Cameoka and Moreira (2021) report three studies with university students who rated photos of White men and of Black men. In Study 1, 75 participants rated, on a scale from 1 (bad) to 10 (good), 10 photos of White men and 10 of Black men with neutral expressions; Study 2 had 40 participants, with faintly smiling faces, and Study 3, 30. The difference between the means of the two groups of photos was significant in Studies 1 and 2 but not in Study 3, which used the photos of White men with the highest ratings and those of Black men with the lowest ratings in the first two studies. The authors take Study 3 as a demonstration of the potential for distortion of a loosely controlled stimulus selection and suggest that stimulus selection can bias measurement. What is measured depends on what was chosen to be measured.

There is also the limit of the fact-checking instrument, which Chapter 2 treats in full: the proportion of false statements among those checked is not the proportion among all statements, and an agency is not a neutral sensor. Mosleh and Rand (2022) use a single fact-checker, PolitiFact, and U.S. elites. The reader should keep in mind, for the remaining chapters, what this section establishes: every falsity rate is a function of who was measured, with what instrument, in what situation, and none of them is, by itself, the rate of a political camp.

1.6 What is known about Brazil

What the verified sources say about Brazil is less than what Section 1.2 says about the United States, and of a different nature. Ricard and Medeiros (2020), in a commentary rather than an empirical study, write that a rhetoric denying the risks of COVID-19 was embraced at the very top of the Brazilian government and propped up by misinformation; and they report that, according to the joint congressional inquiry into fake news (the Comissão Parlamentar Mista de Inquérito, or CPMI das Fake News, set up by Brazil's National Congress in 2019), part of the misinformation about the pandemic attacked decision-makers and public figures, such as state governors, the media, and health experts, in order to delegitimize those who supported social isolation, and that the inquiry had not yet released evidence of direct coordination by

the so-called Gabinete do Ódio ("Office of Hate," a group of presidential aides accused of running online smear campaigns) or by members of the government.

Carvalho et al. (2021) carried out a content analysis, using the IRaMuTeQ software, of posts on Trump's and Bolsonaro's Facebook profiles and of official statements on the websites of the two governments, from the first day on which COVID-19 was reported in each country. The authors point to thematic similarities and to positions contrary to the measures advocated by the WHO. They write that it is not possible to state categorically that there was intentional alignment between the two, but that the analysis shows convergence in disinformation strategies, and that the minimization of the risks of contagion, especially in the Facebook posts, is one of the hallmarks of the similarity. This is an analysis of two senders, not prevalence data.

Recuero et al. (2025) analyzed 811 fact-checks by three Brazilian agencies on the 2018 and 2022 elections, using content analysis and mixed methods, comparing themes, types of disinformation, and the sources that lent legitimacy to the content. They point to a transition from entirely false to misleading content, and from content without sources to content that uses real sources; they read the shift in the platforms of the checked content, from Facebook to WhatsApp and Twitter between 2018 and 2022, as an adaptation of disinformation to more closed spaces, in which counterdiscourses have a harder time circulating. They likewise note that the literature reviewed points to a greater volume of disinformation generated by Bolsonaro supporters in both elections, even though both campaigns appeared simultaneously as sources and targets of disinformation. Lupa, a Brazilian fact-checking agency (Agência Lupa, 2026), describes the same shift in its own database: in 2018, Facebook appeared in 76.5% of its fact-checks; in 2022, WhatsApp was present in 92.4%.

Two limits separate these sources from what Section 1.2 shows for the United States. None of them is a sharing study with a record of behavior, like that of Guess et al. (2019), or a comparison of parties by what their legislators share, like that of Lasser et al.: what exists are fact-checks, which are curated by an agency, and a content analysis of two senders. And the rule of Chapter 2, that the proportion of false statements among those checked is not the proportion among all statements, applies here in full. That Brazilian disinformation has the shape of the American kind, or the reverse, is a hypothesis, and this chapter does not assume it.

1.7 Hate speech: definitions, measures, and data

Among the verified sources, the expression "hate speech" has an explicit definition in Hangartner et al. (2021): hostile language aimed at a group or an individual on the basis of an innate characteristic, real or perceived; the authors selected xenophobic and racist tweets by combining a dictionary, sentiment analysis, and manual annotation. The other studies operationalize it differently, and Table 1.3 brings together what each one measured.

TABLE 1.3

How five verified studies measure hate speech, and what each measure does not allow

Study	Measure of "hate speech"	Sample	Result	What the measure does not allow
Siegel et al. (2021)	Dictionary augmented by machine learning and classification by Reddit communities; the dictionary requires a slur or derogatory term	More than 750 million tweets about the 2016 U.S. election and almost 400 million tweets from a random sample of American users	No persistent increase during the campaign or after the election	Generalizing to other platforms; the definition is conservative
Marino-Fages and Martínez-Pozo (2026, working paper)	BERTimbau classifier and dictionary, by target (homophobia, racism, sexism, insults, political hate)	More than 37.6 million geolocated tweets in Brazil, 2017–2019	Increase after the 2018 election, larger where Bolsonaro received fewer votes, in homophobia and sexism	Measuring the perceived social cost; classifier accuracy is 77% to 79%
Álvarez-Benjumea and Winter (2018)	Hostility rating (7 or higher) given to comments in an experimental forum	180 participants from a crowdsourcing platform	Hostile comments: 5.88% at baseline, 0.99% with moderate removal; counterspeech had no significant effect	Generalizing: paid participants, anonymous forum
Hangartner et al. (2021)	Xenophobic and racist tweets, selected by dictionary, sentiment, and annotation	1,350 Twitter users	Empathy-based counterspeech increased deletion and reduced the creation of xenophobic speech	Effect on those who only read; humor and warnings of consequences without consistent effect

Study	Measure of "hate speech"	Sample	Result	What the measure does not allow
Fazzano and Gallo (2015)	Homophobia as a response class (hitting, insulting, torturing, beating, excluding, discriminating), defined by function	2011 data from Disque 100 (Brazil's federal human rights hotline): 6,809 reports of violations against LGBT people	Records of official complaints; they do not measure the direct prevalence of aggression	Prevalence: these are reports, not the phenomenon

Note. Figures from the articles. Marino-Fages and Martínez-Pozo is a working paper that, according to its cover, has not undergone peer review. The studies are discussed in Chapter 5 (Section 5.5), Chapter 10, and Chapter 12.

Four points organize the table. The first is that the definition decides the number: Siegel et al. say that their dictionary definition is conservative, since it requires the tweet to contain a slur or derogatory term, and Marino-Fages and Martínez-Pozo classify by target and obtain an increase concentrated in homophobia and sexism, with no comparable pattern for general insults, political hate, or political targets. The second is the platform and the cut of the sample: the American study is limited to Twitter and warns that observing a certain kind of speech at one moment does not mean that it has increased, given the bursty character of the platform, on which topics gain prominence for a short time and then return to equilibrium; the Brazilian study uses only tweets posted on Mondays and geolocated in Brazil, and assumes that these two cuts represent the universe of the platform. The third is that "speech" and "act" are different measures, and none of the five studies measures acts against the group (the gap is discussed in Section 5.5). The fourth is that the results neither contradict nor confirm one another: they measure different things in different populations.

On the behavior-analytic side, the literature is scarce. Mizael (2018) found only four behavior-analytic articles on homosexuality in 13 journals and points to functional analyses of prejudiced behaviors, such as jokes maintained by the verbal community, as a gap; and Fazzano and Gallo (2015) define homophobia as a response class of diverse topographies, defined by function rather than by form. This is the definition from which Chapter 5 starts.

1.8 Rival interpretations and limits

The reading proposed in this chapter is that the empirical data reveal magnitudes smaller than the public debate assumes, and that each figure is strictly a function of the instrument used. This perspective faces three rival interpretations. The first holds that partisan asymmetry is an established fact

and that the chapter adopts excessive skepticism; the decisive test between these readings would require a design that kept the total supply of content and actual individual exposure under rigorous control—a caveat that the authors of Garrett and Bond (2021) and Guess et al. (2019) themselves formulate. The second rival thesis argues that, without professional fact-checking, any measurement is unfeasible, which makes the agencies the only thermometer available; the answer of Chapter 2 is that fact-checking measures what was checked, not what circulates. The third is that the definition by intent makes measurement impossible and should be abandoned; the definition by domain, which the studies use, is a pragmatic answer, and it is taken as such here.

The limits of the chapter are those of its sources. Most of the measures come from the United States and from a single platform, over the period of one or two elections; the only large-scale Brazilian database among those on speech is a working paper without peer review; and the prevalence figures for false information are, at best, figures of exposure and sharing, not of effect. Brazil is not compared with the United States, and sharing from a source is not treated as belief or as an act.

Summary

Each definition of false information imposes its own methodological requirements (the sender's intent, the origin of the content, the trustworthiness of the domain, or factual veracity), and research tends to replace the analysis of the statement with the classification of the source because private intent is inaccessible (1.1). The empirical circulation of fake news through individual sharing proved rare in the data of Guess et al., with a higher incidence among older adults, and passive exposure does not equal causal impact (1.2). The partisan asymmetry recorded among legislators in the United States by the criterion of shared domains appears to be a phenomenon peculiar to that institutional context (1.3). The relations between ideological position and disinformation are correlational and depend on the set of items selected and on the supply environment, and a substantial part of the gap in accuracy changes when the contingencies of reinforcement are manipulated (1.4). Self-report in surveys is verbal behavior under the control of the contingencies of data collection, and arbitrary sampling of stimuli can bias the results of measures (1.5). In Brazil, the empirical picture is limited to content analyses, situation reports, and fact-checking agency databases, with no direct behavioral records of sharing on open networks (1.6). Finally, measures of hate speech vary radically depending on the lexical criteria and the collection platform adopted, and they do not capture the emission of concrete acts against the target groups (1.7).

The subsequent chapters draw directly on this framework: Chapters 2 and 3 operate under the principle that every falsity rate is a function of the measuring instrument; Chapter 5 adopts the functional definitions of Section 1.7; Chapter 6 rests on the evidence that sharing represents a tiny fraction of exposure; and Chapter 12 examines the discrepancy between what individuals report verbally and what they actually do when faced with real contingencies.

References

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Cameoka, M. C., & Moreira, M. B. (2021). Preconceito racial: viés na mensuração de atitudes produzido por controle de estímulos [Racial prejudice: Bias in attitude measurement produced by stimulus control]. *Acta Comportamentalia*, *29*(1), 93–112. https://doi.org/10.32870/ac.v29i1.78782

Carvalho, P. R., Sousa, P. C. C., & Schneider, M. A. F. (2021). Desinformação na pandemia de Covid-19: similitudes informacionais entre Trump e Bolsonaro [Disinformation in the Covid-19 pandemic: Informational similarities between Trump and Bolsonaro]. *Em Questão*, *27*(3), 15–41. https://doi.org/10.19132/1808-5245273.15-41

Fazzano, L. H., & Gallo, A. E. (2015). Uma análise da homofobia sob a perspectiva da análise do comportamento [An analysis of homophobia from the perspective of behavior analysis]. *Temas em Psicologia*, *23*(3), 535–545. https://doi.org/10.9788/tp2015.3-02

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), Article eabf1234. https://doi.org/10.1126/sciadv.abf1234

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Hildum, D. C., & Brown, R. W. (1956). Verbal reinforcement and interviewer bias. *The Journal of Abnormal and Social Psychology*, *53*(1), 108–111. https://doi.org/10.1037/h0040518

Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D., & Lewandowsky, S. (2022). Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, *1*(4), Article pgac186. https://doi.org/10.1093/pnasnexus/pgac186

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Marino-Fages, D., & Martínez-Pozo, A. (2026). *Hate in the Tropics: Political leaders and the social acceptability of online hate speech* (Documento de trabajo N.º 391). RedNIE. https://rednie.eco.unc.edu.ar/files/DT/391.pdf

Mizael, T. M. (2018). Perspectivas analítico-comportamentais sobre a homossexualidade: análise da produção científica [Behavior-analytic perspectives on homosexuality: An analysis of the scientific literature]. *Revista Perspectivas*, *9*(1), 15–28. https://doi.org/10.18761/PAC.2017.011

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), Article 7144. https://doi.org/10.1038/s41467-022-34769-6

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, *7*(6), 892–903. https://doi.org/10.1038/s41562-023-01540-w

Recuero, R., Tavares, C., Volcan, T., Pozzebon, M., & Dutra, M. (2025). A desinformação narrada pela checagem: estudo das eleições de 2018 e 2022 [Disinformation as narrated by fact-checking: A study of the 2018 and 2022 elections]. E-Compós, 28, 1–26. https://doi.org/10.30962/ecomps.3039

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. Harvard Kennedy School Misinformation Review, 1(2), Article 013. https://doi.org/10.37016/mr-2020-013

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 U.S. election campaign and its aftermath. *Quarterly Journal of Political Science*, *16*(1), 71–104. https://doi.org/10.1561/100.00019045

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Council of Europe Report DGI[2017]09). Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

instituto WALDEN4 | ACADEMIA WALDEN4



ENCICLOPÉDIA IW4

The concepts in this book, one entry at a time

Reinforcement, stimulus control, equivalence, evidence-based practice: technical entries written clearly, each with its sources at the foot. For studying, reviewing, and citing.

80+ open entries

Free access, no sign-up required

walden4.com.br/enciclopedia

CHAPTER

2

What Fact-Checking Agencies Say (and What They Cannot Say)

In July 2026, Agência Lupa, a Brazilian fact-checking agency, released the report *Anatomia da desinformação eleitoral* [Anatomy of Electoral Disinformation], with the database of its fact-checks over five election cycles, from 2016 to 2024. Among the 23 politicians with more than 30 sentences rated, Jair Bolsonaro had the highest proportion of sentences classified as "False" (Falso): 102 of 236, or 43.2%, in the presidential campaigns of 2018 and 2022. Luiz Inácio Lula da Silva appears in seventh place: 22 of 83 sentences, or 26.5%, in election campaigns he ran in "recently" (Agência Lupa, 2026, pp. 12, 19–20).

On another front, Aos Fatos, another Brazilian fact-checking agency, kept a counter of Bolsonaro's "false or distorted" statements from 2019 to 2022. On December 22, 2022, it reported that he was leaving the presidency with 6,675 (Ribeiro & Menezes, 2022); in July 2025, the same outlet spoke of 6,685 (Menezes & Faustino, 2025). For Lula, Aos Fatos has no equivalent counter: at the end of 2023 and of 2024, it published roundups of the claims he repeated most (Ribeiro, 2023; Menezes, 2024).

Placing 43.2% beside 26.5%, or 6,685 beside a ranking of repeated claims, gives the impression of a comparison. This chapter argues that, with the data published today, neither of the two operations constitutes a legitimate comparison, and it makes explicit what would be needed for it to be one.

This is not skepticism toward the agencies. It is the reading they themselves ask for: Lupa writes, in the same report, that it chose not to produce rankings of who "lied the most," because the volume of fact-checks on each politician depends on public exposure, on the frequency of interviews and debates, on the office held, and on communication style (Agência Lupa, 2026).

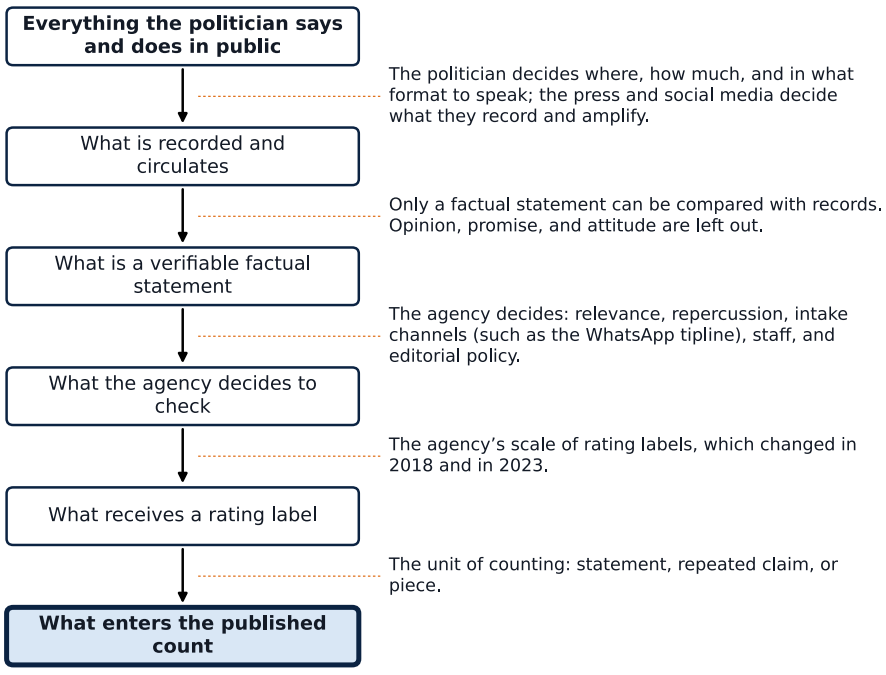
The chapter also serves a function in the book: it sets the rules by which the following chapters read agency data. When a remark appears as "classified as false by agency X," it means exactly that, and nothing more.

2.1 Two series of contingencies

A politician's remark that reaches an agency's count has passed through two series of contingencies. The first is the politician's: where, to whom, how often, and in what format the politician speaks. The second is the agency's: what enters its radar, what it considers relevant and verifiable, what scale it uses to classify, and how it counts. Figure 2.1 summarizes the path. Each arrow is a selection, and only the last stage reaches the reader.

FIGURE 2.1

From statement to classification: where selection happens



Only the last stage reaches the reader: the published proportions are fractions of what passed through the earlier ones.

Note. Original diagram by the author. It does not represent proportions: the stages have no known size. The mechanisms noted on the arrows come from Lupa itself (Agência Lupa, 2026) and, for the unit of counting, from the sources in Table 2.1.

Three points in the figure deserve detail, because the agencies themselves document them.

First, what is checked depends on how much public exposure the politician has. Lupa observes that candidates who speak more and are more incisive produce more verifiable claims and may therefore "have more 'False' ratings and wrongly stand out as the biggest liar" (Agência Lupa, 2026, p. 19). The Aos Fatos counter shows the same effect on the scale of a single day: one interview, on the Flow Podcast on August 8, 2022, yielded at least 26 false or distorted statements in five hours and 21 minutes, and the record for the term was January 15, 2021, with 61 remarks checked (Santos, 2022). Both cases suggest that the count tracks, in part, the time of exposure.

Second, what enters the agency's radar depends on its intake channels. Lupa maintains a WhatsApp tipline and takes part in Meta's fact-checking programs, which, in its words, tends to concentrate part of the flow of content received on Meta platforms, so that the results about platforms should be read as "a portrait of the universe monitored by Lupa" (Agência Lupa, 2026, p. 82). And the report's database reflects only what the agency published in its Elections section, conditioned by editorial choices and by criteria of journalistic relevance.

Third, selection can be read as behavior of the agency itself. Kniess et al. (2024) start from a scenario in which journalism faces a crisis of legitimacy and trust, and they ask to what extent two agencies linked to traditional media companies, Estadão Verifica and Fato ou Fake, seek to reinforce their own epistemic authority, that is, their credibility as a source of what is true, through fact-checking. To answer, they collected all the texts the two agencies published with Bolsonaro's name in the title between January and May 2022 and with Lula's between January and May 2023, 119 fact-checks in all (95 from Estadão Verifica and 24 from Fato ou Fake), and submitted this corpus to content analysis. The agencies concentrated their coverage on similar subjects under the two governments, and their verification emphasized sources from journalism itself, that is, their own materials or those of other traditional media companies. According to the authors, this suggests limits to the transparency of the verification, because much of the information used has already passed through editorial filters that are not always clear to the public.

In behavior-analytic language, fact-checking is the fact-checker's verbal behavior. It is under the control of antecedent stimuli (the politician's statement and the audience), but also of institutional contingencies of reinforcement (such as editorial policy) and of consequences that go beyond the statement checked. This does not make it arbitrary. It makes the count dependent on a

selection process that has to be described in order to be read. That these consequences include, for example, credibility with the audience is a hypothesis compatible with Kniess et al. (2024), not a datum of this chapter.

2.2 What each source counts

Before comparing numbers, ask of each one what was counted. Box 2.1 brings together the five questions, and Table 2.1 answers them for the sources used in this chapter.

BOX 2.1

Five questions before comparing counts

1. What is the *unit*: sentence, statement, repeated claim, piece, or mention?
2. What is the *denominator*, that is, what was checked, and is it published?
3. What are the *period* and the *situation*: campaign, term in office, calendar year?
4. What is the classification *scale*, and has it changed?
5. *Who chose* what went in, and by what rule?

TABLE 2.1

What each source counts and what it publishes as a denominator

Source	Unit counted	What was published	Denominator published	Period and situation
Lupa, report (Agência Lupa, 2026)	Checked sentence (unit of analysis within a fact-check)	"False" sentences per politician, among those with more than 30 sentences	Yes, per politician: 236 (Bolsonaro) and 83 (Lula)	Election campaigns from 2016 to 2024; Bolsonaro, 2018 and 2022; Lula, cycles not specified
Aos Fatos, Bolsonaro counter (Santos, 2022; Ribeiro & Menezes, 2022; Menezes & Faustino, 2025)	"False or distorted" statement ("inaccurate" and "incorrect" in 2025)	6,042 (Aug. 2022); 6,675 (Dec. 2022); 6,685 (Jul. 2025)	No	Term in office, Jan. 2019 to Dec. 2022
Aos Fatos, Lula roundups (Ribeiro, 2023; Menezes, 2024)	Repeated claim	Ranking of the most repeated, with the number of occasions (21 times in 2023; at least 24 in 2024)	No	Calendar years 2023 and 2024; term in progress

Source	Unit counted	What was published	Denominator published	Period and situation
Aos Fatos, ten-year review (Menezes & Faustino, 2025)	Published fact-checks; mentions in disinformation pieces	More than 19,000 fact-checks; about 15,000 statements by politicians and officials; mentions: Lula, 677, and Bolsonaro, 566	Overall total, not broken down by politician or by term	2015 to 2025
Kniess et al. (2024)	Fact-check (piece) with the president's name in the title	119 fact-checks: 95 from Estadão Verifica and 24 from Fato ou Fake	Yes, for the corpus	Jan.–May 2022 (Bolsonaro) and Jan.–May 2023 (Lula)
Recuero et al. (2025)	Fact-check (piece)	811 fact-checks by three agencies	Yes, for the corpus	2018 and 2022 elections

Note. All sources are fact-checks published by the agencies themselves or studies about them; in all of them, the agency is the one that chooses what goes in (question 5).

The table makes evident what public debate tends to hide. Only Lupa publishes a numerator and a denominator broken down by politician—restricted, however, to campaign periods, not to terms in office. Aos Fatos, for its part, releases for Bolsonaro a numerator with no denominator; and, for Lula, it reports repeated claims, which constitute a different unit of analysis. Its overall totals (more than 19,000 fact-checks; about 15,000 statements by politicians and officials and almost 4,000 disinformation pieces since 2015) are not broken down by politician or by term (Menezes & Faustino, 2025) and do not serve as a denominator for either of the two.

2.3 Why the proportion among checked statements is not the proportion among statements made

The fraction published by an agency is a proportion among statements that have passed through the selection in Figure 2.1. Let p be the proportion of false factual statements in what the politician says; sF, the probability that a false statement is checked; and sT, the probability that a true one is. The observed proportion among checked statements is

$$p \cdot s\mathrm{F} \div (p \cdot s\mathrm{F} + (1 - p) \cdot s\mathrm{T}),$$

which coincides with p only when sF = sT—that is, under the assumption that the agency checks false and true statements with identical probability. None of the published reports authorizes this assumption of equality, nor does any of them allow the selection ratio $r = s\mathrm{F} \div s\mathrm{T}$ to be estimated. Lupa reports that

it also checks in order to confirm what is true ("a central (and frequent) function of fact-checking") and that 34.4% of the sentences checked were rated "True" (Agência Lupa, 2026, pp. 12, 29), which shows that the selection is not only of suspect statements, but it does not say with what relative probability.

Table 2.2 shows, with hypothetical numbers, what the ratio *r* does to the observed proportion.

TABLE 2.2

Hypothetical illustration: the same formula with different selection ratios

Actual falsity rate (p)	Selection ratio (r = sF ÷ sT)	Observed proportion among checked statements
20%	1.0	20.0%
20%	1.5	27.3%
20%	2.0	33.3%
20%	3.0	42.9%
10%	4.0	30.8%

Note. Hypothetical values, chosen to illustrate the formula; they are neither data nor estimates about any politician. The observed proportion is p·r ÷ (p·r + 1 – p).

The last row is the point. A politician with half the actual falsity rate (10% versus 20%) appears with a higher observed proportion (30.8%) than another with an actual rate of 20% and a ratio of 1.5 (27.3%), if selection favors checking what is false more strongly in the first case. The same reasoning applies to discrepancies between different agencies or between different time intervals. The conclusion is not that these data are invalid, but that the difference between observed proportions combines two independent processes—the repertoire emitted by the politician and the agency's selection filter—variables that the published databases do not allow one to isolate.

There is also a second source of imprecision. Sentences from the same debate, interview, or televised questioning are not independent: they share topic, audience, and moment. Lupa notes that, in its fact-checks of debates and interviews, a single report may analyze dozens of candidates' sentences (Agência Lupa, 2026). The Flow Podcast case (Section 2.1) is the same thing seen from the side of Aos Fatos.

2.4 What Lupa allows one to say

Lupa is the only source examined that publishes a numerator and a denominator per politician, with the same scale for all. It is, for this reason, the material closest to a comparison. Table 2.3 brings together what the report says about the two presidents, and Figure 2.2 places them among the ten politicians in the published ranking.

TABLE 2.3

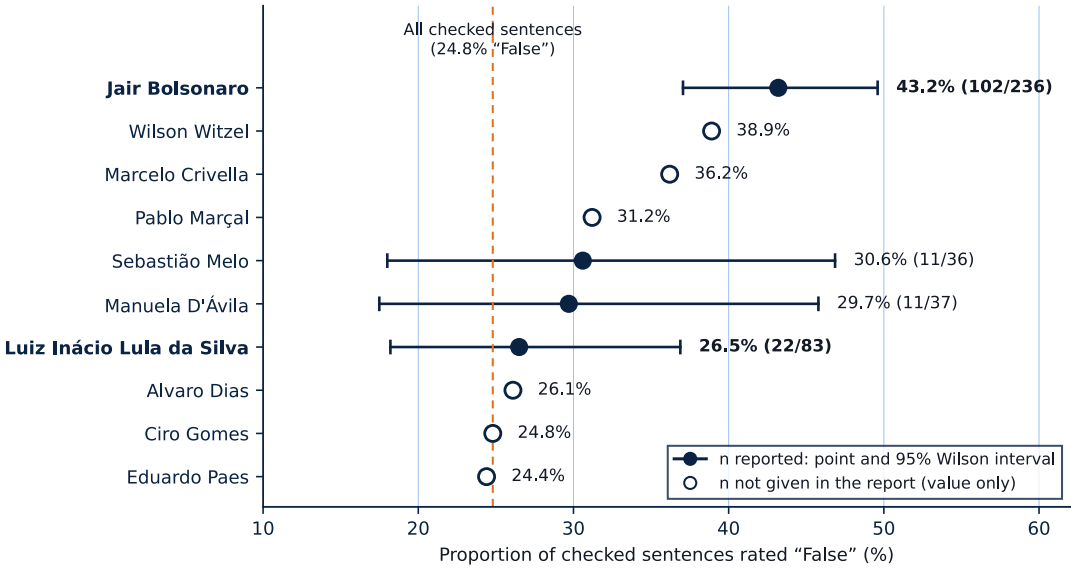
What Lupa's report says about Bolsonaro and Lula

	Jair Bolsonaro	Luiz Inácio Lula da Silva
Cycles considered	Presidential campaigns of 2018 and 2022	"Election campaigns he ran in recently"; cycles not specified
Sentences checked	236	83
"False" sentences	102	22
Proportion "False"	43.2%	26.5%
Rank among the 23 politicians with more than 30 sentences	1st	7th
95% Wilson interval, calculated by the author	37.1% to 49.6%	18.2% to 36.9%
Pattern of the falsehoods, according to Lupa's linguistic analysis	Incorrect numbers; absolute and misleading claims; "never," "not ever," "none," "always"	Claims of authorship of public policies, or of having been first to adopt them

Note. Source: Agência Lupa (2026). The Wilson interval sets the 95% confidence margin for a proportion under the theoretical assumption that the checked sentences are an independent random sample of the repertoire emitted by the speaker. Because the checked items share common contexts and result from deliberate editorial filters (Section 2.3), this condition of independence is not strictly met; the interval serves here only as an estimate of the order of magnitude of the imprecision due to sample size, not as a confirmatory hypothesis test.

FIGURE 2.2

Proportion of sentences classified as "False" among those checked by Lupa, by politician



Note. Values from the chart on p. 20 of Lupa's report, which considers only actors with more than 30 checked sentences (Agência Lupa, 2026); the number of sentences comes from the text and is available for only four politicians. The order is the report's and does not measure "who lies the most" (Section 2.3). The politicians belong to different parties, offices, and levels of government. 95% Wilson intervals calculated by the author.

Four observations keep Table 2.3 from turning into a hasty conclusion.

It is a sentence, not a remark or a piece. The report's unit is the sentence checked within a fact-check. The database has 632 publications, 397 fact-checks (checagens) and 235 verifications (verificações), which generated 3,366 units of analysis, 3,080 from fact-checks and 286 from verifications (Agência Lupa, 2026).

It is a campaign, not a term in office. The 236 and the 83 come from election campaigns. For Bolsonaro, the cycles are those of 2018 and 2022, in which he was a candidate and, in the second, an incumbent president running for reelection. For Lula, the report speaks of campaigns he ran in "recently" and does not specify them. The speaker's situation (candidate, president, opposition) changes what is said and what is checked, and the table does not control for it.

"False" is one label among several. Across the whole database, "False" totals 765 records (24.8%); "Exaggerated" (Exagerado), 502 (16.3%); and "True, But… / Lacks Context" (Verdadeiro, Mas… / Falta Contexto), 434 (14.1%) (Agência Lupa, 2026). The last two together add up to 30.4%, more than "False," and Lupa observes that, in four of the five cycles, they exceeded "False" (p. 27). The

published report does not break down the distribution of the other labels by politician. A comparison made only with "False" leaves out most of what the agency classifies as problematic overall.

The scale changed. Lupa started with eight labels; it added "Underestimated" (Subestimado) in 2018; and, in 2023, "Keep an Eye On" (De olho), "Too Early to Tell" (Ainda é cedo), and "True, But..." gave way to "Lacks Context" (Agência Lupa, 2026, p. 28). For their analyses, Lupa's own researchers merged "True, But..." and "Lacks Context." Any series that crosses 2023 must do the same.

DATA	In Lupa's database, 43.2% of Bolsonaro's 236 checked sentences in the 2018 and 2022 presidential campaigns were classified as "False," compared with 26.5% of Lula's 83 checked sentences in campaigns he ran in recently. The proportion of "False" in the database as a whole is 24.8%.
INTERPRETATION	Among the sentences Lupa chose to check in campaigns, the proportion of "False" was higher for Bolsonaro than for Lula and than the database mean. The Wilson intervals (37.1% to 49.6% and 18.2% to 36.9%) almost touch; they indicate considerable imprecision, especially for Lula.
HYPOTHESIS	Part of the difference may reflect speaking styles (Lupa describes numbers and absolute claims in one case and authorship and pioneering in the other), the speaker's situation in each cycle, and the selection ratio r of Section 2.3. None of the three was tested here.

2.5 What Aos Fatos allows one to say

Aos Fatos publishes two kinds of counts about the presidents, and each has its own unit and criterion: a continuous counter, kept between 2019 and 2022, and annual roundups on Lula. The two passages below say what each allows one to conclude.

The counter. Between 2019 and 2022, Aos Fatos kept a counter of the president's false claims, which monitored "interviews, live streams, speeches, and posts since the inauguration, in January 2019" (Santos, 2022). On August 17, 2022, it stood at no fewer than 6,042; on December 22, at the end of the term, at 6,675, "an average of 4.6 per day" (Ribeiro & Menezes, 2022); in July 2025, a text by the same outlet speaks of 6,685, sometimes as "false or inaccurate" statements, sometimes as "incorrect" ones (Menezes & Faustino, 2025). In the texts

published by the outlet, there is no explicit justification for the discrepancy of ten records, nor any specification of the criterion adopted for including the label "inaccurate."

Two things, however, are clear. The first is that the counter publishes a numerator without a denominator: the texts released by the outlet do not give the total number of the president's statements that were checked, so that 6,675 does not express the proportion of any set. The second is that the unit, the statement, is cumulative and tracks exposure, as the Flow Podcast case showed. The average of 4.6 per day is consistent with 1,461 days in office (6,675 ÷ 1,461 = 4.57, the author's calculation), but it is an average per day in office, not per remark or per hour of exposure.

The Lula roundups. In 2023, the roundup was carried out "based on the president's speeches, interviews, and live broadcasts" (Ribeiro, 2023). In 2024, Aos Fatos "analyzed all the speeches, addresses, and interviews" by Lula and set aside the checkable statements; over the course of the year, it checked the most repeated ones, and, at the end of December, those classified as "False" or "Not Quite So" (Não é bem assim) were counted to form a ranking (Menezes, 2024).

The unit, therefore, is the repeated claim, and the ranking answers a different question from the counter's: "which falsehoods did he repeat most?", not "how many false statements did he make?". In 2023, the most repeated was the claim that he had inherited 14,000 stalled public works projects (21 times; classified as false) (Ribeiro, 2023). In 2024, it was the claim that the PT (Workers' Party) government ended hunger in 2014 (at least 24 occasions; classified as false), and another, repeated on at least 17 occasions, was that he was the only president of the Republic without a university degree (classified as false) (Menezes, 2024). For Bolsonaro, the ten-year review reports that his most repeated claim was that there had not been a single corruption allegation in his government (Menezes & Faustino, 2025).

The two sets would become comparable only if both presidents were subjected to both questions: the ranking of repeated claims and the total number of statements checked, in the same period and by the same criterion.

DATA	Aos Fatos published 6,675 "false or distorted" statements by Bolsonaro during his term (Dec. 2022) and 6,685 in 2025; for Lula, it published annual rankings of repeated claims classified as false. Neither gives the total number of statements checked.

INTERPRETATION	These are faithful descriptions of each series' count, but of different units (statement and repeated claim) and without a denominator. They do not allow one to say who "said more falsehoods" or who said them in a greater proportion.
HYPOTHESIS	That Bolsonaro's counter grew at a pace similar to his exposure (long interviews, live streams) is plausible given the Flow Podcast case, but it has not been tested.

2.6 Remarks and mentions: whose falsehood is it?

A frequent unit error is to treat two kinds of pieces as equivalent: the fact-check of what the politician said and the verification of content that circulates about the politician. Lupa separates them. A "fact-check" (checagem) is the verification of specific remarks, statements, or claims, usually attributed to public figures; a "verification" (verificação) is that of viral or potentially viral content with no identified author (Agência Lupa, 2026).

Aos Fatos reports, in its ten-year review, that Lula and Bolsonaro were the most fact-checked public figures and also the most cited in disinformation pieces: 677 mentions for Lula and 566 for Bolsonaro. And it adds that, "while the PT leader is attacked more by misleading pieces, Bolsonaro, for the most part, has his actions inflated" (Menezes & Faustino, 2025). Kniess et al. (2024) found that more than half of the 119 fact-checks in their corpus were aimed at refuting rumors unfavorable to Lula.

These numbers describe what circulates *about* each politician, not what each one said. Counting disinformation pieces that cite a politician as if they were that politician's remarks is a unit error. In the following chapters, only remarks attributed to the politicians themselves are used, and content about them is treated as a different object.

2.7 Do the agencies agree with one another?

A fact-check is a classification, and classifications can diverge. There are three sources of variation.

The first is the scale: Lupa's changed twice (Section 2.4). The second is the content: Recuero et al. (2025), in 811 fact-checks by three Brazilian agencies in the 2018 and 2022 elections, found a transition from completely false to misleading content and from content without sources to content that uses real sources. What is classified changed, and with it the distribution of labels. The

third is agreement between the agencies themselves: there are empirical studies on consistency between fact-checkers in election debates (e.g., Lelo, 2023), but their specific data are not incorporated into this account (see Appendix D).

As long as the third question cannot be answered with data, each classification is treated as attributed to the agency that made it, not as an established fact.

2.8 What fact-checking cannot reach

Fact-checking operates on factual statements, which are compared with records: official data, documents, recordings. Behavior analysis offers a distinction that is useful for knowing where this ends.

CONCEPT IN USE

Belief and attitude as verbal behavior (Guerin, 1994)

Guerin treats attitudes and beliefs as verbal behavior shaped by verbal communities, not as reports of an internal state. A *belief* ("Nuclear reactors are unsafe") can be disputed with evidence or by the opposite statement. The corresponding *attitude* ("I do not like the use of nuclear energy") cannot be tackled in the same way: the speaker can reply, in Guerin's formulation, that, "despite all you've said," he or she still does not like it. For this reason, presenting a topic as an attitude rather than as a belief is often reinforced, and the difference between the two forms is a contingency of the community of listeners, not a property of the statement (Guerin, 1994).

CONCEPT IN USE

Autoclitic (Guerin, 1994)

Autoclitics are qualifications added to a statement that modify its effect upon the listener, such as "It seems to me that ...," "I had always believed that ...," or "I like dogs somewhat" said to a cat-lovers society (Guerin, 1994; Skinner's 1957 term, cited by Guerin). They can reduce negative consequences and allow what was said to be requalified after a negative reply.

What the agencies assess is restricted to factual statements—equivalent, in Guerin's (1994) behavior-analytic vocabulary, to the class of beliefs, which maintain correspondence with verifiable antecedent stimuli. By contrast, statements formulated as attitudes (such as "this government is the worst in history") or mediated by autoclitic qualifiers ("in my view," "I'm just asking") are predominantly under the control of the audience's immediate social consequences and fall outside the scope of factual veracity labels. Moreover,

Guerin (1994) shows that statements in the form of an attitude often operate as mands: the boss who remarks in public that she does not like employees who drink coffee while they work functionally exerts control over the listener's behavior so that the listener stops doing it, regardless of any private intent.

The consequence for reading the agencies is direct. The label measures the correspondence between what was said and the records. It does not measure the function of what was said for the audience. Two statements, one checkable and the other an attitude or an autoclitic qualification, can have similar effects on those who hear them, and only the first enters the count. Chapters 3 and 5 analyze function; this chapter limits itself to noting that the part of discourse that migrates from the factual form to the opinion form moves out of the reach of fact-checking without its function changing.

DATA	Guerin (1994) describes that beliefs can be disputed with evidence and attitudes cannot, and that presentation in the form of an attitude is often reinforced. Lupa and Aos Fatos check factual statements.
INTERPRETATION	A statement that moves from the factual form to the opinion form ceases to be checkable, even if what it does to the listener is similar.
HYPOTHESIS	That part of this migration is a response to the contingencies of fact-checking itself (avoidance of countercontrol, Chapter 6) is a hypothesis to be tested with dated remarks, before and after fact-checks; there are no data on this in this chapter.

2.9 Protocol for a comparable sample

To compare the two presidents without the choice of cases depending on the result, the rule below was fixed before any data collection.

1. **Agency.** A single one, Lupa, which has a single scale and publishes a denominator. Mixing agencies would mix scales and selection criteria.
2. **Windows.** The first 45 months of each term: from January 1, 2019, to September 30, 2022 (Bolsonaro) and from January 1, 2023, to September 30, 2026 (Lula), by publication date of the fact-check. Both have the same duration and end before the first round of voting (October 2, 2022, and October 4, 2026).

3. **Unit.** The sentence attributed to the politician in a fact-check. Viral content about the politician and remarks by third parties are left out. When a piece contains several sentences, each one is counted; when the markup does not separate them, the piece is counted and flagged.
4. **Scale.** The original labels are kept. "True, But..." (before 2023) and "Lacks Context" (after) are grouped, as Lupa does; "Keep an Eye On" and "Too Early to Tell" are kept apart, as "no verdict."
5. **Outputs.** Number of sentences per label and proportion, with a Wilson interval, for each president; a separate table for single-sentence and multiple-sentence pieces.
6. **Sensitivity.** Repeat excluding the official campaign periods, excluding debates and long interviews, and by quarter, to see whether the difference appears or disappears.
7. **Validation.** Compare the counts for the 2018 and 2022 campaigns with the report's 102 of 236, and Lula's with the 22 of 83, to estimate the coverage of the data collection.
8. **Limits declared in advance.** The selection is the agency's; there is no denominator of total remarks; the exposure of the two varies over the term; the data collection depends on the coverage of the source.

The data collection was not carried out, for three reasons. (a) The robots.txt files of Lupa and Aos Fatos restrict automated collection by computational tools, and these parameters were strictly respected. (b) The public ClaimReview repository of Google Data Commons, which aggregates structured fact-checks with this metadata, has irregular historical coverage of Lupa: it records 3,369 fact-checks concentrated between 2022 and September 2024 (with only 12 entries in 2019, 3 in 2020, and 4 in 2021, and no records under the current domain), which makes it impossible to define symmetrical time windows. (c) Lupa's report does not specify Lula's cycles. The paths still open are: Lupa sharing the database of its Observatory (the database behind the report exists), or the data collection being carried out by the author, under his own responsibility.

Summary

A fraction published by an agency is the proportion among the statements that have passed through two series of selection, the politician's and the agency's (Figure 2.1). For this reason, before comparing numbers, one should ask of each what the unit is, what the denominator is, what the period and the situation are, what the scale is, and who selected what went in (Box 2.1).

Once these questions are asked, the data published today do not support a comparison between Bolsonaro and Lula. Lupa publishes a numerator and a denominator per politician (Bolsonaro, 102 of 236, or 43.2%; Lula, 22 of 83, or 26.5%), but these are fractions from campaigns, with different and unspecified cycles, and the label "False" is one among several. Aos Fatos publishes, for Bolsonaro, a numerator without a denominator (6,675 in 2022; 6,685 in 2025) and, for Lula, rankings of repeated claims: these are different units. And pieces about a politician, the mentions, are not that politician's remarks.

Fact-checking, finally, applies strictly to factual statements (beliefs, in Guerin's 1994 conceptualization). Attitude formulations and autoclitic modulations remain immune to verification, even though they may exert analogous behavioral functions on the audience. A legitimate comparison between Bolsonaro and Lula would require scrutiny by a single agency, bounded by symmetrical time windows within the terms in office and accompanied by the respective denominator. Because such a body of evidence does not exist in current publications, the methodological rule for its eventual construction is set out in Section 2.9.

References

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst, 17*(1), 155–163. https://doi.org/10.1007/bf03392661

Kniess, A. B., Comel, N., & Marques, F. P. J. (2024). Checagem de fatos e autoridade jornalística no Brasil: uma análise do Fato ou Fake e do Estadão Verifica [Fact-checking and journalistic authority in Brazil: An analysis of Fato ou Fake and Estadão Verifica]. *Comunicação e Sociedade*, *46*, Article e024020. https://doi.org/10.17231/comsoc.46(2024).5665

Lelo, T. (2023). Assessing the consistency of fact-checking in political debates. *Journal of Communication, 73*(6), 587–600. https://doi.org/10.1093/joc/jqad028

Menezes, L. F. (2024, December 23). *No segundo ano de governo, Lula desinformou mais sobre fome e educação* [In his second year in government, Lula spread the most disinformation about hunger and education]. Aos Fatos. https://www.aosfatos.org/noticias/lula-2024-desinformou-fome-educacao/

Menezes, L. F., & Faustino, M. (2025, July 7). *Como a desinformação transformou o debate político brasileiro nos últimos dez anos* [How disinformation transformed Brazilian political debate over the last ten years]. Aos Fatos. https://www.aosfatos.org/noticias/desinformacao-transformou-debate-politico-dez-anos/

Recuero, R., Tavares, C., Volcan, T., Pozzebon, M., & Dutra, M. (2025). A desinformação narrada pela checagem: estudo das eleições de 2018 e 2022 [Disinformation as narrated by fact-checking: A study of the 2018 and 2022 elections]. E-Compós, 28, 1–26. https://doi.org/10.30962/ecomps.3039

Ribeiro, A. (2023, December 27). *Em primeiro ano de mandato, Lula repete alegações enganosas sobre fome, economia e governo Bolsonaro* [In his first year in office, Lula repeats misleading claims about hunger, the economy, and the Bolsonaro government]. Aos Fatos. https://www.aosfatos.org/noticias/lula-alegacoes-enganosas-2023/

Ribeiro, A., & Menezes, L. F. (2022, December 22). *Uma retrospectiva das mentiras do governo Bolsonaro* [A retrospective of the Bolsonaro government's lies] [Comic-strip story]. Aos Fatos. https://www.aosfatos.org/noticias/retrospectiva-mentiras-governo-bolsonaro/

Santos, L. F. dos. (2022, August 17). *Bolsonaro supera 6.000 declarações falsas ou distorcidas no início da campanha pela reeleição* [Bolsonaro surpasses 6,000 false or distorted statements at the start of his reelection campaign]. Aos Fatos. https://www.aosfatos.org/noticias/bolsonaro-supera-6000-declaracoes-falsas-ou-distorcidas-no-inicio-da-campanha-pela-reeleicao/

Skinner, B. F. (1957). Verbal behavior. Appleton-Century-Crofts.



PART II

Operant Analysis of the Discourses

3 False Statements as Verbal Operants

4 Political Attitudes and Prejudice as Equivalence Classes

5 Hate Speech: Disguised Mands and Aversive Control

6 Digital Bubbles as Verbal Communities and the Avoidance of Countercontrol

7 Voting as Verbal Behavior

CHAPTER

3

False Statements as Verbal Operants

On August 28, 2022, in a presidential debate broadcast by Band, Folha, TV Cultura, and UOL, Luiz Inácio Lula da Silva said: "The lowest deforestation in the Amazon happened under my government." Agência Lupa, a Brazilian fact-checking agency, classified the statement as "False": the lowest deforestation rate in the Legal Amazon recorded by the Prodes system of Inpe (Brazil's National Institute for Space Research) was that of 2012, under the government of Dilma Rousseff, at 4,600 km^2, the lowest in the series begun in 1989 (Agência Lupa, 2026). In the same debate, Lula said he was proud to have appointed "the first Black man to the Supreme Court"; Lupa also classified this as "False," because before Joaquim Barbosa, appointed in 2003, two other Black justices had served on the Supremo Tribunal Federal (STF, Brazil's Supreme Federal Court), Pedro Lessa and Hermenegildo Rodrigues de Barros (Agência Lupa, 2026). On September 29, 2022, in a debate on TV Globo, Jair Bolsonaro said: "No country in the world bought vaccines in 2020"; "False," according to Lupa, which cites a survey by Poder360, a Brazilian news outlet, according to which at least 50 countries began vaccinating in 2020. And on August 3, 2018, in an interview on GloboNews's Central das Eleições, he said: "I have never been homophobic"; "False," in Lupa's fact-check, which recalls the conviction for collective moral damages upheld by the Rio de Janeiro State Court of Justice in November 2017, for statements the agency describes as homophobic in content.

The four statements share a feature that is not their falsity. All have the grammatical form of someone stating a fact, and all contain one of the terms that Lupa itself lists among those that make a claim absolute: "lowest," "first," "no," "never" (p. 41). The question that usually follows is whether the speakers were lying, believed what they said, or were mistaken. This chapter changes the question. Instead of inferring what the speaker had in mind, it asks which

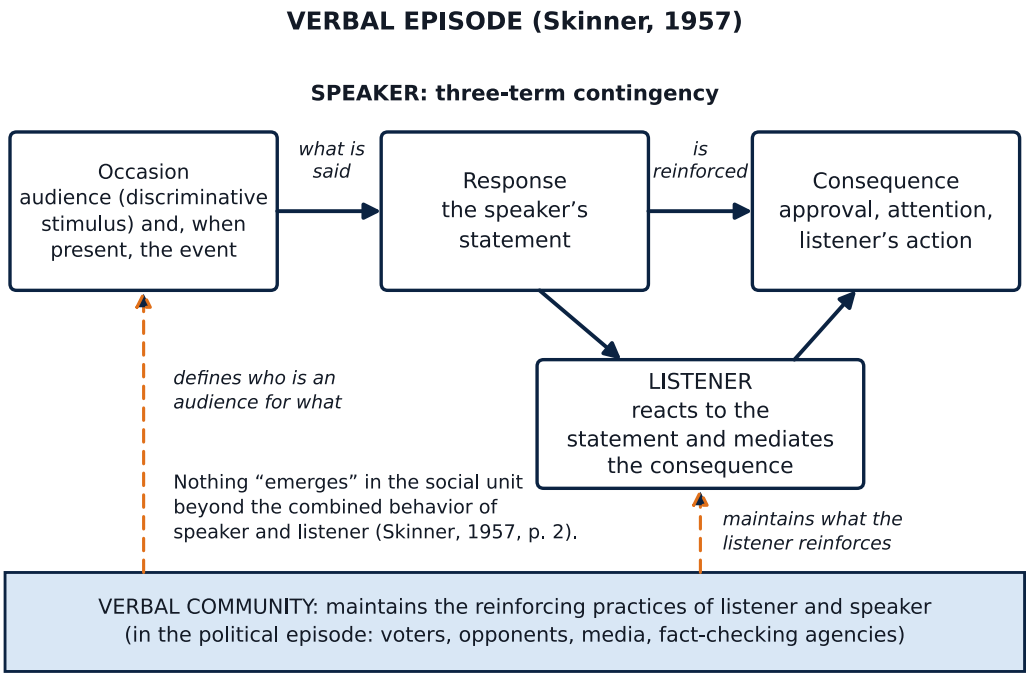
variables control the emission of such a statement and what its grammatical form conceals. It is the core chapter of the book: it presents the analysis form that the following chapters use and applies it to these four cases. They were not chosen at random, nor are they a sample: they are the examples that Lupa itself reproduced in its report to illustrate each politician's style (Agência Lupa, 2026), and they are included here because they are documented in a source that was read, with date, context, and the classification attributed to whoever made it.

3.1 The politician's speech as verbal behavior

Skinner (1957) defines verbal behavior as a kind of behavior necessarily reinforced through the mediation of other people. The definition has practical consequences. What a politician says does not change the world directly: it changes what listeners do, and it is through what they do that the speaker's behavior is consequated. Skinner calls the behaviors of speaker and listener taken together a total speech episode, and adds that "there is nothing in such an episode which is more than the combined behavior of two or more individuals. Nothing 'emerges' in the social unit" (p. 2). Figure 3.1 brings the scheme together.

FIGURE 3.1

The verbal episode: the speaker's three-term contingency, mediated by the listener and maintained by the verbal community



Note. Original diagram by the author, based on Skinner (1957). It does not represent data. In the occasion, the audience enters as a discriminative stimulus.

Three consequences of this definition guide the rest of the chapter. The first is that grammatical form does not identify the operant. Skinner notes that terms such as "declaration" or "asserts," applied to the tact, refer to additional processes, the autoclitics, and that the tact, considered on its own, does not on that account amount to an assertion (Skinner, 1957). A sentence in the indicative, with subject and predicate, is merely a sentence in the indicative, and only the history and the variables present tell whether it was controlled by an event, by other words, or by what the listener would do next. The second is that a politician's speech never has a single audience. Skinner discusses multiple audiences and states that the same verbal response may produce different effects depending on the audience in whose presence it is emitted. When each audience is presented separately, control over verbal behavior is much more precise. On the other hand, when distinct audiences are present at the same time, the simultaneous production of conflicting consequences is common. In a televised debate there are the voters the candidate wants to persuade, those who already support him, the opponents on stage and, later, those who fact-check. These different listeners constitute conflicting audiences.

CONCEPT IN USE

Three-term contingency, function × topography, and the verbal episode (Skinner, 1957)

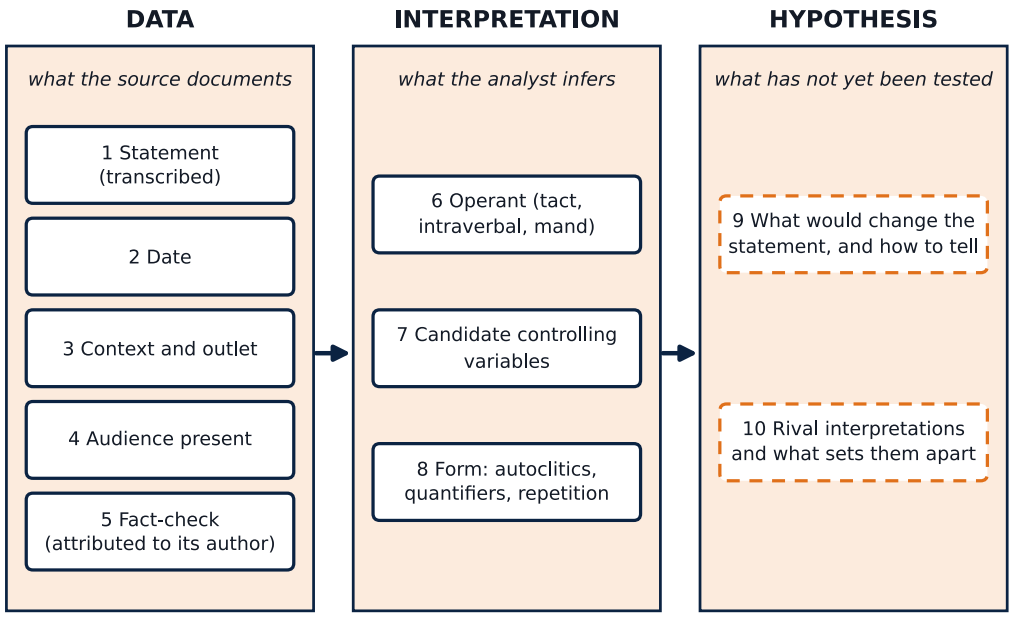
A verbal operant is defined by what controls the response, not by its form. For Skinner (1957), the form of a response is not enough to classify it as a mand, and, in general, identifying any verbal operant requires knowing what kind of variables the response is a function of. The same response topography, that is, the same form (e.g., "the vaccine is dangerous"), can therefore be a tact, an intraverbal, or a mand, depending on the occasion, the history, and the consequence. In other words, what classifies it are the controlling variables of the verbal response "the vaccine is dangerous." The *three-term contingency* is the unit: occasion (discriminative stimulus), response, consequence. In verbal behavior, the occasion includes the audience, a discriminative stimulus in whose presence verbal behavior is typically reinforced, and the consequence comes from the listener. The *verbal episode* includes the interactions between speakers and listeners whose verbal responses exert multiple stimulus functions on each other's behavior.

3.2 The analysis form

Every analysis in this book goes through the same form, and the form exists to force a separation between what is known and what is inferred. Figure 3.2 shows its fields. The first five are data: the statement, transcribed as the source gives it; the date; the context and the outlet; the audience present, when the source documents it; and the fact-check, always with the name of whoever made the classification, because "false" is not a property of the sentence but the decision of an agency with a method (Chapter 2). The next three are interpretation: the probable operant, the candidate controlling variables, and the additional verbal elements (autoclitics, quantifiers, repetition). The last two are hypothesis: what would change the statement and how one would know, and the rival interpretations together with what separates them. Where the source does not document something, the form says "not documented" and does not fill in the field by inference.

FIGURE 3.2

The analysis form: what is data, what is interpretation, and what is hypothesis



In each empty field, the form says "not documented"; it does not fill it in by inference.

Dashed outline: what, in the chapter's cases, is hypothesis and not result.

Note. Original diagram by the author. The version for use, with blank fields, is in Appendix A.

A reading rule accompanies the form, and it is Skinner's own. Faced with records of a speaker's behavior in which the conditions of emission are unknown, one can reconstruct a probable verbal history, but only rarely can inferences about the relevant variables be directly checked. The criterion for understanding a causal analysis of verbal behavior is the ability to predict the occurrence of instances and, eventually, to produce or control them. The cases in this chapter are records of statements, made by third parties, without access to the conditions under which they were made; by this criterion, their analyses are mere interpretations. That is what is worth calling interpretation: a plausible reading, which says which variables could control the verbal response and what one would need to observe in order to decide among them.

Table 3.1 brings together the operants that the form uses, with the controlling variable of each (Skinner, 1957) and an example constructed for this chapter, which comes from no source and serves only to fix the difference.

TABLE 3.1

Verbal operants: what controls the response and a constructed example from political speech

Operant	What controls the response (Skinner, 1957)	Constructed example
Mand	Deprivation or aversive stimulation of the speaker: "the response is reinforced by a characteristic consequence and is therefore under the functional control of relevant conditions of deprivation or aversive stimulation"; "the response has no specified relation to a prior stimulus" (pp. 35–36)	"Vote for me."; "Enough stalling."; "Let's change Brazil"
Tact	Object, event, or property of an object or event: "a response of given form is evoked (or at least strengthened) by a particular object or event or property of an object or event"; "in the presence of the object or event a response of that form is characteristically reinforced in a given verbal community" (pp. 81–82)	"The candidate shoved his opponent during the debate," said in the presence of what happened
Echoic	Auditory verbal stimulus: "the response generates a sound-pattern similar to that of the stimulus" (p. 55). There is phonetic correspondence between stimulus and vocal responses	Repeating ("The only good criminal is a dead criminal"), in reply, the exact expression used by the opponent ("The only good criminal is a dead criminal")

Operant	What controls the response (Skinner, 1957)	Constructed example
Textual	Nonauditory verbal stimulus (text): "a vocal response is under the control of a nonauditory verbal stimulus" (p. 66). Emission of vocal responses with phonemes corresponding to written stimuli	Reading aloud a passage of the bill on the floor of the legislature
Transcription	Verbal stimulus, with "a point-to-point correspondence between different dimensional systems" (p. 71). Copying: emitting a written response with letters corresponding to the written stimulus. Dictation: emitting a written response with letters corresponding to the phonemes in the vocal stimulus	Copying: typing on one's phone a tag matching the one generated by an AI on the computer. Taking dictation: taking down an opponent's statements.
Intraverbal	Verbal stimulus without point-to-point correspondence with the response: "some verbal responses show no point-to-point correspondence with the verbal stimuli which evoke them" (pp. 71–72)	Responding "will never be defeated" to the opening of the chant "The people united"

Note. The definitions are from Skinner (1957), on the pages indicated; the examples column is by the author. Some categories were supplemented by the author, since Skinner's most precise definition is given in fragments spread across paragraphs. The six types are operants defined by controlling variables, not by forms: the same sentence may belong to more than one type on different occasions.

3.3 Tact or intraverbal? The continuum and the tact form

Only for the tact does it make sense to assess whether the "verbal response" (spoken, written, or gestured) is true or not, since its controlling variable is a nonverbal stimulus, that is, events in the world. Skinner defines it as "a verbal operant in which a response of given form is evoked (or at least strengthened) by a particular object or event or property of an object or event," adding that "in the presence of the object or event a response of that form is characteristically reinforced in a given verbal community" (pp. 81–82). The verbal community establishes it because tacts increase listeners' contact with the environment. When someone says that a candidate for governor hit another with a chair during a debate, that person is emitting a tact, provided the chair attack actually took place. At the same time, a listener who did not watch the debate will have access to what happened (more specifically, will have access to the report of what happened) through the speaker's emission of the tact. In the mand, in turn, the listener has access to what may reinforce the speaker's verbal response, since the mand is the only verbal operant whose responses specify the reinforcer and, in many cases, the listener's action.

CONCEPT IN USE

Generalized social reinforcement (Skinner, 1957; Guerin, 1994)

Any event that usually precedes many different reinforcers can act as a reinforcer for behavior under any relevant state of deprivation or aversive stimulation; a common example is approval, which can amount to no more than a nod or a smile from someone who usually provides a range of reinforcers (Skinner, 1957). Because this reinforcer does not depend on a particular state of deprivation, the form of the response comes to depend on the stimulus present: this is how the tact is established, and this is why generalized reinforcement is, for Skinner, the key to practical and scientific discourse. For Guerin (1994), attitudes are statements maintained by this generalized social reinforcement from a verbal community. Because they precede several other reinforcers available for many operants, generalized reinforcers bear no relation of correspondence to specific response topographies; that is, they reinforce an endless multiplicity of responses. Approval, illustrated by the remark "Well done!," will reinforce different responses, such as solving a math equation or pronouncing a word correctly in German, for example.

The purity of the tact, however, is a matter of degree. Skinner calls a tact maintained by fully generalized reinforcement, which allows exclusive control by the nonverbal antecedent stimulus, "pure" or "objective," and this rarely happens. Ideally, scientific discourse is concerned with creating conditions (scientific method and rigor) for scientific tacts to approach this ideal. Even in scientific discourse, distortions occur, since scientific communities reinforce certain reports of research findings more than others. Not to mention the absence of epistemological consensus. In the political sphere, in turn, distortions are much more likely, since we are referring to control over the behavior of other people, in this case, the listeners.

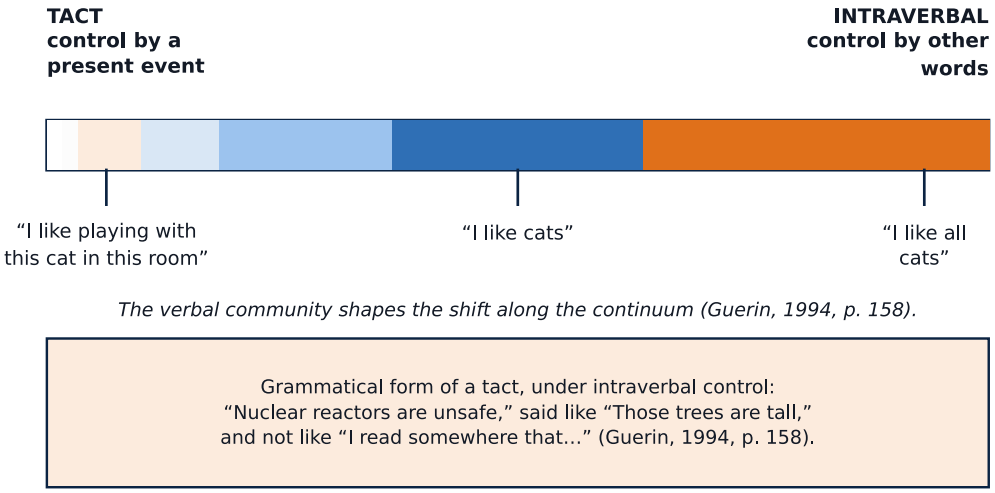
The intraverbal, although reinforced by the same reinforcer as the tact, is controlled by verbal stimuli that bear no point-to-point correspondence between stimulus and response. The vast majority of the content taught in formal education is intraverbal. Taking history as an example, we learn about King João VI's raid on the Brazilian coffers when he returned to Portugal by reading or hearing about this historical event, since we did not witness it. What we say or write about this event is considered intraverbal responding. Guerin (1994) carries this idea over to the behavior of reporting beliefs and attitudes. For him, distinguishing between tacts and intraverbals is particularly challenging in many concrete examples, something behavior analysis has not yet satisfactorily clarified. Verbal communities shape the shift along a continuum, from "I like playing with this cat in this room," which has some plausibility as a tact,

to “I like cats” and “I like all cats,” which are intraverbals. Many attitudes concern events the person has never experienced, so what is known comes from newspapers and from other people. Figure 3.3 summarizes.

What makes the intraverbal relevant to those who study disinformation is what the community reinforces. For Guerin (1994), verbal communities reinforce the emission of intraverbals as if they were tacts (pp. 155, 158): saying “Nuclear reactors are unsafe” in the same grammatical form as “Those trees are tall” makes a more powerful and influential statement than “I read somewhere that nuclear reactors are unsafe” (p. 158). Guerin (1992) analyzes socially constructed knowledge as intraverbals, tacts with loose discrimination, and verbal behavior under powerful social controls from the group. The same sentence, without autoclitics that add information about its controlling variables (e.g., “people are saying that…”), can be either an intraverbal or a tact, depending on the controlling variable. Tacts, in general, carry more credibility among listeners than intraverbals. The sentence “I saw a certain married politician going into a brothel” (i.e., the autoclitic “I saw” specifies the verbal response as a tact) carries more credibility than “I was told that a certain married politician went into a brothel” (i.e., the autoclitic “I was told” specifies the response as an intraverbal). On the other hand, the sentence “A certain married politician went into a brothel” can be either a tact or an intraverbal because it contains no autoclitic elements that inform the listener of its controlling variable.

FIGURE 3.3

The tact–intraverbal continuum and the tact form



Note. Diagram by the author, with Guerin’s (1994) examples. Guerin gives “I like cats” and “I like all cats” as intraverbals; the intermediate position is illustrative. It does not represent data.

An example shows what this means for the analysis. Mota et al. (2023) studied ten statements by Bolsonaro about the CoronaVac vaccine, between July 2020 and January 2021. Using search engines, they selected the ten statements with the greatest repercussion that mentioned the vaccines directly, and recorded for each one its number, date, context (outlet and occasion), and transcribed content. They classified the information disorders using Wardle and Derakhshan's typology, in which disinformation is false information deliberately created to cause harm and misinformation is false information without intent to cause harm; malinformation is the presentation of true information, but out of context, in order to cause harm. The most frequent type was misinformation (47.1%), followed by disinformation (35.3%) and malinformation (17.6%). In one of the statements, on October 29, 2020, during his weekly livestream, Bolsonaro said that the vaccine that took the least time to be made available to the population had taken four years. The authors read the sentence as partially true information (in fact, the fastest-developed vaccine up to then, against mumps, had taken four years) used in a false context, with the intention, according to them, of casting doubt on the COVID-19 vaccines, whose development took less time. What the analysis form adds to the authors' reading is the separation between what the record shows and what they inferred. The record shows a sentence that, taken alone, corresponds to a verifiable fact. It does not show whether the response was controlled by the fact (in which case the failure would lie in the context in which it was used) or by a sequence of words learned as an argument (in which case the fact would be tangential). Saying "false context," as Mota et al. do, is describing the effect of the sentence on the listener, not the variable that controlled it.

3.4 The distorted tact: when the listener's consequence displaces control by the event

Skinner deals with the distortion of the tact in the chapter on the special conditions affecting stimulus control (1957). The starting point is that generalized reinforcement is the key to practical and scientific discourse: it places the speaker's behavior under the control of the current environment and lets the listener respond to the speaker's words in place of direct contact with the situation. When the correspondence with the stimulating situation is sharply maintained, the response is called objective, valid, true, or correct. But stimulus control is never perfect, and it may be distorted by special consequences provided by a given listener or by listeners as a whole, in which case the response is called subjective, prejudiced, or wishful. When generalized reinforcement is contingent on specific verbal response topographies, whether or not they correspond to the stimuli they refer to, distorted tacts become more

frequent (Ribeiro, 1989). If a person says that a candidate for governor is corrupt, even without any evidence for it, that verbal response has probably been reinforced with approval and support by the candidate's opponents.

Skinner's description is graded, and the gradient matters. In the minor case, the speaker inflates the size of the fish he caught or plays down the danger of an attack; the response is still controlled, in part, by the event. In a greater case, control is invented, and the response is emitted when the event has not occurred: the child who lost a penny, said so, and was given one starts saying it when no penny has been lost, and the boy who cried "Wolf!" is the stereotype. The clearest example of the mechanism is the troubadour. He begins by recounting actual exploits; certain parts receive special approval because they interest or flatter his listeners; these parts survive in later tellings, and, as that differential reinforcement continues, he starts to stretch the account, inflating the scale of the battle and the heroism of those involved, until he severs all stimulus control, "describing" scenes he never witnessed. Distortion is thus a product of selection by consequences: the listener reinforces some versions, the speaker repeats them, and, step by step, correspondence ceases to be what decides.

CONCEPT IN USE

Distorted tact and verbal subcommunity (Skinner, 1957; Guerin, 1992)

A *distorted tact* is a tact whose control by the event has been displaced by special consequences supplied by the listener. Skinner (1957, p. 149) describes degrees: the speaker "stretches the facts"; stimulus control is "invented," and the response occurs when the circumstances in which it is typically reinforced are absent; in the extreme, the response is emitted under circumstances that control an incompatible response, and this is called a lie. For Guerin (1992), control over tacting may lie in the hands of a smaller *verbal subcommunity*: if a small group reinforces certain reports and punishes others, the characteristics of the group come to control what is reported, even if the consequences are generalized, and Skinner called these tacts distorted.

Skinner also describes a mixture that occurs frequently, the impure tact. When a housewife says "Dinner is ready" chiefly because it brings her listeners to the table, the response functions almost like the mand "Come to dinner!"; this blend of the controlling relations typical of the tact and of the mand is what he calls an impure tact. Figure 3.4 shows the logic. The point is that a response can be a function of both variables at the same time, and the sources of strength add up. This matters for the interpretation of statements: even

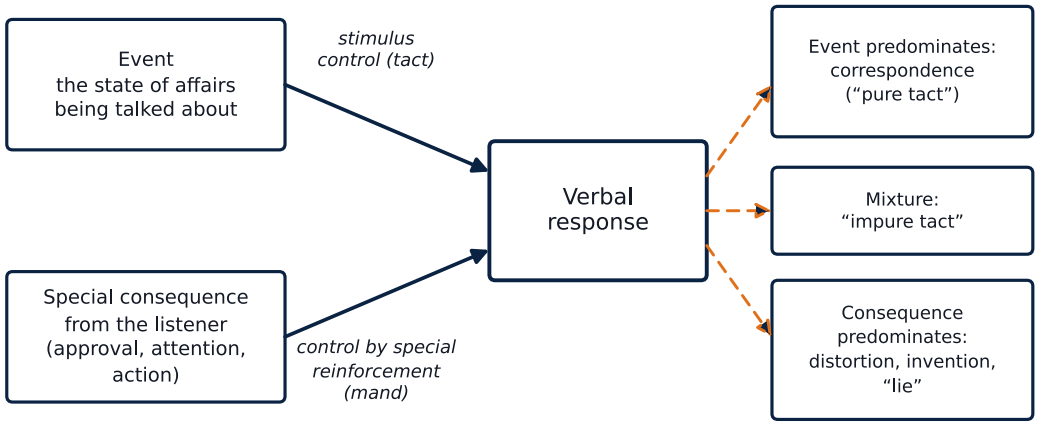
when the event controls part of the response, the listener's consequence may control another part, and falsity, measured by correspondence with the event, does not tell what the proportion is.

Skinner adds two predictions of interest to politics. The first concerns the cost to the system. When distortion leads the listener to react ineffectively, the speaker–listener system may deteriorate: the listener may withhold reinforcement or punish the speaker's responses. The distorted tact is "temporarily more effective" because it exploits a stronger disposition of the listener to respond; the community stops giving the child a penny and may punish him for lying. The second concerns the condition for stability: the system remains stable only when, for the listener, correspondence with the controlling stimuli has no practical importance, as in literature. And there is the observation about aversion: a tact's stimulus control is particularly prone to distortion when the response serves to avoid or escape aversive consequences; the suborned witness is the example.

These predictions were written for a dyad (one speaker, one listener) and for a community in which the listener reacts to speech with actions that depend on its corresponding to the facts. Political speech differs on three points, and it is on these points that the analysis advances by hypothesis. First, there are several audiences, and whoever reinforces the distortion may not be whoever punishes it (Section 3.1); the deterioration predicted for the dyad may not occur as long as the speaker's behavior is reinforced by one audience and punished only by another, which has no access to him. Second, for part of the audience, correspondence with the facts may be, in Skinner's terminology, of no practical importance, and the function of the sentence becomes something else: the effects it produces on listeners. Guerin (1994) notes that reporting attitudes can work as a social ritual, in which the reinforcement lies in the exchange itself. Third, when the speaker belongs to a subcommunity that reinforces certain responses and punishes others, what one has is what Guerin (1992) describes: what is reported comes to depend on the group. Chapters 6 and 9 deal with the last two.

FIGURE 3.4

Multiple control: the event and the listener's special consequence act on the same response



Note. Diagram by the author, based on Skinner (1957). The three outcomes are types, not proportions; no source measures how much of the strength of a political response comes from each variable.

DATA	Skinner (1957) describes, without measuring, three properties of the distorted tact: it is a function of special consequences from the listener; it is temporarily more effective than the tact; and it leads the speaker–listener system to deteriorate, except when correspondence does not matter to the listener. Guerin (1992) describes the shift of control to small groups.
INTERPRETATION	Political speech can be read by the same logic, with the caveats of multiple audiences and noninformative functions. Statements classified as false by an agency are candidates for an analysis of distortion, not demonstrated cases of it.
HYPOTHESIS	In a field in which correspondence with the facts is of little practical importance to the reinforcing audience, the distorted tact does not deteriorate and tends to stabilize. No verified source measures this for politicians' statements.

3.5 Repetition: the strength of the repeated claim

Repetition appears in two positions, and they should not be confused. There is the speaker's repetition, when the same sentence is said again on different occasions (which makes it an intraverbal response), and the repetition the listener undergoes, when encountering the same sentence again. The first is a

dependent variable in operant analysis: frequency of emission is a measure of response strength. The second is an independent variable in much of the psychology of disinformation.

On the speaker's side, Skinner offers three points of support. The first is the observation that a story or joke that went over especially well tends to be told again (1957, p. 148): special measures of reinforcement "tell the speaker what is worth talking about" (p. 148). The second is the warning that repetition may indicate no more than an excess of response strength, rather than a specific controlling variable. The third is that, when verbal behavior fails to change the conditions that give it strength, one response may give way to another, above all when straight repetition is punished, and what appears is "repetition with variation" (p. 228). The prediction is testable: if straight repetition is punished, the form should vary. This happens often in writing, where teachers punish sentences repeated throughout a text.

What the verified sources document for Brazil is the number of times certain claims were repeated. Aos Fatos, a Brazilian fact-checking agency, recorded that the claim that the governments of the PT (Workers' Party) had ended hunger in Brazil was made in Lula's inauguration speech, from the parlatorium of the Planalto Palace, the presidential office, and repeated on at least seven other occasions in 2023, and that it had been recurrent in his speeches since the campaign (Ribeiro, 2023); in 2024, it was the most repeated claim of the year, on at least 24 occasions, and was classified by Aos Fatos as false for comparing two different indicators (Menezes, 2024). The claim that he would be the only president without a university degree was repeated on at least 17 occasions in 2024 and classified as false (Menezes, 2024). In the other camp, the equivalent source is different in nature. Mota et al. (2023) transcribe two statements by Bolsonaro, from December 17, 2020, and January 15, 2021, that repeat, in very similar words, two things: that he had already caught the virus and had antibodies ("why take the vaccine again?," "why take the vaccine?") and that the laboratories, in their contracts, do not take responsibility for side effects; and, according to the authors, who cite a January 2021 report in the newspaper O Globo, he made at least ten public criticisms of CoronaVac between July 2020 and January 2021.

Four observations separate what this shows from what it does not. The first: the two sources measure different things (Aos Fatos counts occasions on which a claim was repeated in a year, based on speeches, interviews, and livestreams in 2023 and, in 2024, on all of the president's speeches, addresses, and interviews; the authors transcribe ten statements chosen for their repercussion), and so the counts are not comparable. The second: repetition of the claim about hunger continued in 2024 after the 2023 survey had already classi-

fied it as false (Ribeiro, 2023; Menezes, 2024); this is a datum that does not say that Lula knew of the classification, nor that the audience knew of it. The third: what is recorded is the emission, not what maintains the emission. The fourth: the repetitions recorded here are, on the surface, straight repetition, and the chapter does not know whether on other occasions the form varied.

On the listener's side, the verified source is a set of experiments. Effron and Raj (2020) brought together four experiments and a pilot, two of them preregistered, with 2,587 participants in total, to ask whether encountering a false headline again changes moral judgment about spreading it. In Experiment 1, 138 Americans from an online panel saw six of 12 false headlines four times, in a familiarization phase in which they rated, each time, an aspect unrelated to falsity; they then judged all 12, six already seen and six new, having been told that nonpartisan fact-checkers had confirmed that those events did not happen. Having seen a headline once or four times made participants judge it less unethical to publish and share it on seeing it again, even when it was labeled false and they did not believe it, and even after statistically controlling for how likable and popular it seemed; and perceiving the headline as less unethical predicted a greater inclination to express approval of it online. In an experimental setting, people shared repeated headlines more than new ones (Effron & Raj, 2020). The authors speculate that repetition may reduce moral condemnation by making the information seem intuitively true, and they discuss other mechanisms. In operant terms, the result is compatible with one reading: repetition changes what the audience does with the sentence (it condemns less, shares more), that is, it changes the consequence the speaker receives for repeating. This is the chapter's reading, not the authors' conclusion. And the design (headlines, experimental participants, a simulated environment) is far from a campaign speech.

DATA	A claim was repeated on at least 24 occasions in 2024 by Lula (Menezes, 2024), after being classified as false in 2023 (Ribeiro, 2023); two things said by Bolsonaro on 12/17/2020 reappear on 1/15/2021 (Mota et al., 2023). In experiments, repetition reduced moral condemnation and increased sharing (Effron & Raj, 2020).
INTERPRETATION	Repetition is a measure of response strength, and strength is an effect of reinforcement history; the experimental literature suggests that the audience is more tolerant of what it has already heard. The records do not identify which consequences maintained the repetitions.

HYPOTHESIS Repetition by politicians is maintained, in part, by the audience's reaction to it, and it should vary in form when straight repetition comes to be punished. Testing this requires a record of the audience's reactions on each occasion and of the form of the claim before and after the fact-check.

3.6 Autoclitics of source and of certainty: what is added and what is removed

In Chapter 2, the autoclitic appeared as a qualification of the statement that modifies its effect on the listener (Guerin, 1994). Skinner (1957, ch. 12) analyzes it in full, and three families are of interest here.

The first family is that of source autoclitics. Skinner notes that asking "Did you see it, or did someone tell you?" amounts to asking whether the response was a tact or was echoic or intraverbal to another person's verbal behavior, and that in many communities courtesy calls for opening one's speech with an autoclitic that shows where what is said comes from; an autoclitic may also indicate that what is said is already accepted by the listeners or by people in general. They are the counterweight to what Section 3.3 described: the community reinforces intraverbals in the form of tacts, and the source autoclitic is precisely what would say that the response is not a tact. Figure 3.5 puts forward the hypothesis that, along a chain of retransmission, source and weakness autoclitics are lost and what remains is the tact form; Skinner describes how an atypical response in a conversation tends to be taken up and handed on from one speaker to the next, but no verified source follows a chain of this kind.

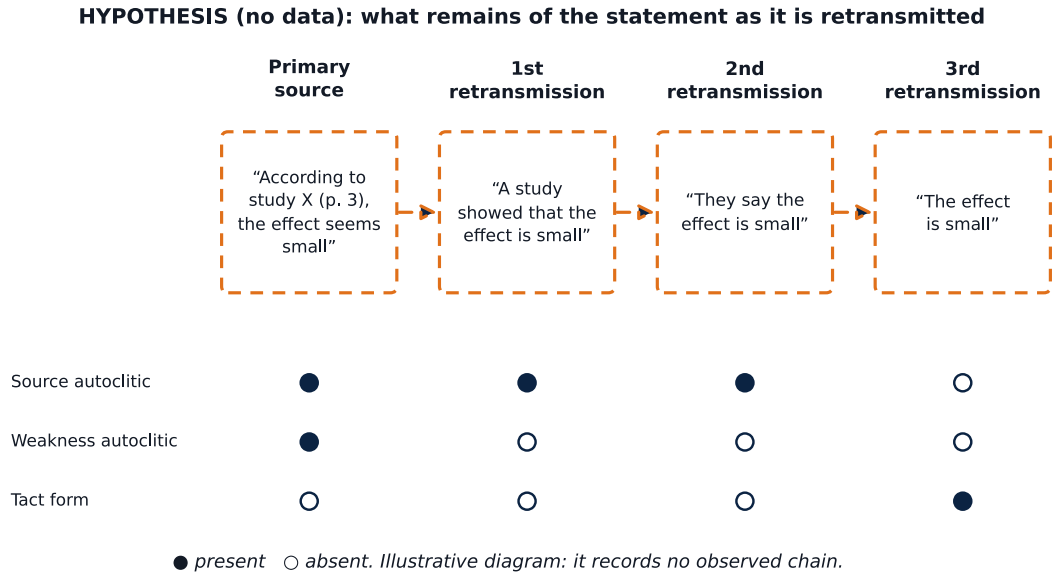
CONCEPT IN USE

Autoclitics of source and of certainty (Skinner, 1957)

The autoclitic is verbal behavior based on other verbal behavior of the same speaker, which modifies the listener's reaction to it. *Descriptive* autoclitics inform the listener of the kind of operant they accompany ("I see that...," for what has been read; "I remember") and of the strength of the response: "I guess" and "I believe" indicate insufficient stimulation or a poorly conditioned response (p. 315); "I know," "I assure you," and "I swear" suggest the strength of the response; they could be dropped without altering the kind of effect produced, but they make it more precise. There are those that indicate the speaker's emotional condition ("I am happy to say"; p. 316) and those that point to the source ("They say..."). *Qualifying* autoclitics include "no," which has the force of a mand and may be roughly translated "Don't act upon this response as an unextended tact" (p. 324), and the assertive autoclitic, which enjoins the listener to accept a state of affairs. *Quantifying* autoclitics ("all," "some," "no") modify the listener's reaction to the whole sentence.

FIGURE 3.5

Hypothesis: loss of source and weakness autoclitics along a chain of retransmission



Note. Hypothesis proposed here. The sentences are constructed; the diagram records no observed chain and serves to state what would need to be observed: the statement at each link and the autoclitics of each.

The second family is that of weakness and certainty autoclitics, and an example appears in one of the statements recorded by Mota et al. (2023). On December 24, 2020, Bolsonaro said, in his weekly livestream: "The efficacy of that vaccine in São Paulo, it seems it's way down there, right? I'm not going to

release a percentage here, because if I get it wrong by 0.001% I'll get beaten up by the media, but it seems the percentage is way down there" (p. 320). The authors classify as false the claim that the percentage was "way down there" (p. 323). The sentence contains "it seems" twice, which is a weakness autoclitic in Skinner's sense (it indicates a response based on insufficient stimulation), and an explicit reference to the audience that can punish it: the media. The form, therefore, informs the listener of the degree of strength of the response, and the reference to the media describes the contingency that, according to the speaker himself, would control him if he gave the number. The statement does not have the tact form that Section 3.3 described: it presents itself as a guess. In the hypothesis of Figure 3.5, it is a statement like this, with a weakness autoclitic, that would lose the autoclitic when retransmitted; Mota et al.'s record does not show whether this happened.

The third family is the most pertinent to Lupa's data. Absolute terms ("always," "never," "all," "none," "only," "first," "largest," "lowest") and strong negations ("does not exist," "there is no") are, in Skinner's classification, quantifiers and qualifiers: "all" is equivalent to "always it is possible to say" and "no" to "it is never possible to say" (p. 329), and "no" has the force of a mand. They define the extent to which the listener is asked to accept the response, and Skinner notes, regarding the assertive autoclitic (and not the quantifiers), that any accompanying condition that would weaken the listener's reaction, such as a contradiction from someone else or dubious circumstances, leads the speaker to intensify it.

Lupa measured the frequency of these patterns in the statements it fact-checked during election periods, and the result deserves careful reading. According to the report, claims with absolute terms were classified as "False" in 40.6% of the fact-checks in election periods, and sentences with negations of the type "does not exist" or "there is no," in 45.9% (Agência Lupa, 2026); those mentioning the electoral process, in 33.2%; those attacking opponents, in 27.9%; and those containing numbers, percentages, or amounts, in 18.3%, although these more often receive "Exaggerated," "Underestimated," or "True, but / lacks context" (p. 50). Of the 936 fact-checks classified as "Exaggerated" or "True, but / lacks context," 737 (78.7%) contained some marker of scale, comparison, quantity, performance, authorship, or legacy. The text does not report the numerators and denominators of the percentages of 40.6%, 45.9%, 33.2%, and 27.9%, nor whether the base of each percentage is the set of fact-checks of each pattern.

There are at least three readings, which are not mutually exclusive. The logical reading: a universal or negative sentence is refuted by a single counterexample, and a hedged sentence is not; the rate of "False" would be, in part, a

property of the logical form, and not of the speaker. The selection reading: agencies check what is verifiable and what circulates; Lupa acknowledges that the volume of fact-checks of each politician depends on time in the spotlight, on interviews and debates and, "above all," on communication style, and that the data reflect only what Lupa published in its Elections section, conditioned by editorial and relevance choices. And the operant reading, which is proposed here: the absolute autoclitic is emitted where the listener's reaction needs to be strengthened, that is, on occasions of dispute, contestation, or doubt. The third can only be tested with a record of the occasion, which the report does not provide.

3.7 The statement as a rule for the audience

A political statement can have another function besides describing a state of affairs: describing a contingency, that is, telling the listener what will happen if he or she does this or that. "Whoever takes this vaccine will turn into an alligator" describes an event (the effect of the vaccine) but, above all, a relation between the listener's behavior and a consequence. In the behavior-analytic literature, this operant and the behavior it evokes are treated as rules, and rule-governed behavior is the listener's behavior with respect to them.

CONCEPT IN USE

Rule and rule-governed behavior (Baum, 1995; Cerutti, 1989)

For Baum (1995), a rule is a verbal discriminative stimulus produced by the speaker under the control of a contingency, that is, of a relation between behavior and its consequences. Cerutti (1989) defines rule-governed behavior as the combination, in complex instructions, of elementary discriminations established beforehand. Behavior is said to be rule-governed when its probability of occurrence is shown to be affected by contact with a rule. The property of interest here is *insensitivity to contingencies*: instructed performances are less sensitive to changes in the programmed consequences (Hayes et al., 1986).

The experimental literature available here, all of it from the laboratory, is consistent with the idea that the rule, by itself, can control the listener's behavior without the listener contacting the contingency. In Galizio (1979), six students aged 18 to 25, recruited through a part-time job advertisement in which pay depended on performance, turned a lever to avoid losses of money, signaled by a red light and a tone; four labeled amber lights indicated which component of the multiple avoidance schedule was in effect. Accurate instructions controlled performance in each component, and instruction following was eliminated when inaccurate instructions led participants to contact a contin-

gency of monetary loss; while the programmed loss was withdrawn, participants followed the inaccurate instruction, with no tendency for rates to decline. Hayes et al. (1986) gave participants a task in which pressing buttons moved a light through a maze, and moving the light earned points worth chances at cash prizes; movement followed a multiple schedule of fixed ratio 18 and differential reinforcement of low rate 6 seconds, with components alternating every 2 minutes. In the second experiment, lights paired with the instructions "Go Fast" and "Go Slow" were presented and later withdrawn. The authors point out that instructed performances are less sensitive to changes in consequences and propose that part of this insensitivity is due to additional contingencies introduced by instructional control, such as the social consequences of following the rule.

For the analysis of statements, the consequence is twofold. The first is that the falsity of a rule may lie in the description of the contingency, and not in the description of the event: "if you turn into an alligator, that's your problem," recorded by Mota et al. (2023) in the statement of December 17, 2020, has a descriptive part (the side effect) and a rule part (the consequence for whoever gets vaccinated). The authors read it as a probable satirical or hyperbolic device, to be considered alongside the potential impact of a statement by a head of state. On the analysis form, what is recorded is the grammatical form (a rule, with a described consequence) and what is hypothesized is the function: if the listener who follows it contacts the real contingency, Galizio's experiment suggests that rule following is extinguished; if the listener does not contact it, or if following the rule has social consequences, the insensitivity described by Hayes and colleagues makes rule following resistant. The second is dependence on the audience: the false rule persists to the extent that the listener does not come into contact with the real contingency and that other contingencies, the social ones, maintain rule following. No verified source measures this with political statements.

3.8 Cases: the same four statements, by the same method

Table 3.2 applies the form to the four statements from the opening. Two are by Lula and two by Bolsonaro, all from the examples that Lupa reproduces in its report, all classified by it as "False," under the same method and in the same publication. For each one, the form repeats what the source documents and what it does not. The operant column is interpretation, in the terms of Section 3.2.

TABLE 3.2

Four statements classified as "False" by Lupa: data, interpretation, and what the source does not document

Case	Statement, date, context, and classification (data)	Interpretation (operant)	Not documented
1. Lula	"The lowest deforestation in the Amazon happened under my government." Band, Folha, TV Cultura, and UOL debate, 8/28/2022. Lupa: "False"; the lowest rate recorded (4,600 km², 2012) was under the government of Dilma Rousseff (PT)	Tact about his own government, with a superlative; distortion in the direction of credit, or an intraverbal from a narrative about "the PT governments"	The question or argument he was responding to; audience reaction; whether the figure had already been fact-checked
2. Lula	"I am proud to have appointed the first Black man to the Supreme Court." Same debate. Lupa: "False"; two Black justices had served on the STF before the appointment of Joaquim Barbosa in 2003	Emotional autoclitic ("I am proud") plus a tact with an ordinal; a historical fact, usually acquired as an intraverbal, emitted in the tact form	Same; whether there was a question on the topic; reaction
3. Bolsonaro	"No country in the world bought vaccines in 2020." TV Globo debate, 9/29/2022. Lupa: "False"; at least 50 countries began vaccinating in 2020 (Poder360); in Brazil, vaccination began on 1/17/2021	Tact with a negative universal quantifier; a possible impure-tact function, depending on what the listener would do with the sentence (hypothesis; the occasion is not documented)	The preceding question or argument; whether the sentence was recurrent; reaction
4. Bolsonaro	"I have never been homophobic." Interview on Central das Eleições (GloboNews), 8/3/2018. Lupa: "False"; conviction for collective moral damages upheld by the TJRJ (Rio de Janeiro State Court of Justice) in Nov. 2017, for statements made in 2011 on the TV program CQC	Tact of his own verbal behavior, with "never"; possible distortion under an aversive contingency (hypothesis)	The question that occasioned it; whether the sentence was repeated; reaction; response to corrections

Note. Agência Lupa (2026). The statements, the classification, and the justifications appear in images of the fact-checks reproduced in the report (clippings in verificacao/evidencias/); the interpretation column is by the author. "False" is the label Lupa applies to information that is demonstrably incorrect.

Lula's cases have in common the first person and credit: "under my government," "I am proud to have appointed." The report presents them as instances of a style, and says that a significant share of Lula's "False" statements during campaigns relate to authorship of or pioneering in public policies, programs, laws, and institutional milestones (Agência Lupa, 2026). For the analysis form, what matters is that, in case 1, Lupa's justification points to the true figure (the lowest rate in the series) under a government of the same party, so that the statement diverges from the figure in a single variable: to whom it is attributed. There are at least two readings. In one, the listener's consequence (approval for credit) displaced control: this is the tact distorted in the direction of credit. In the other, the speaker emitted an intraverbal, learned within a narrative about "the PT governments," and the first person is merely the form it took at that moment. They are distinguished by two observations the report does not provide: what the same speaker says about the same figure before audiences that do not reinforce credit, and whether the form changed after the fact-check. In case 2, the sentence has two parts: Lupa's justification acknowledges Lula's appointment of Joaquim Barbosa in 2003 and points to two Black predecessors on the STF, so that the failure lies in "first." "I am proud to" is, in Skinner's vocabulary, an autoclitic that indicates the speaker's emotional condition and emphasizes the listener's personal relation to him; the ordinal is the point of failure.

Bolsonaro's cases, in the report, illustrate another style: Lupa says he tends to receive "False" when he uses numbers, percentages, and statistics, and when he makes categorical claims and accusations against opponents, with "never," "not ever," "none," and "always" appearing frequently. Both cases have an absolute quantifier. In case 3, the counterexample is public and verifiable (the United Kingdom began vaccinating on December 8, 2020, according to Lupa), which is exactly what makes the universal form risky: a single exception refutes it. In case 4, the sentence is a tact of one's own verbal behavior, and Skinner (1957) writes that the community sets up the contingencies required for self-descriptive behavior when it has some reason to ask "What did you say?" and "Why did you say that?" (p. 314). Guerin (1994) adds that a question can mand the reply. The source does not give the question, and it is the question that decides whether the response is emitted under the contingency of an accusation, which is the condition in which, according to Skinner, the tact is especially likely to be distorted. The analysis form leaves the field blank, and the reading as distortion under an aversive contingency remains a hypothesis.

There are other statements by Bolsonaro documented with date, context, and transcription among the ten that Mota et al. (2023) analyzed, and they were used in Sections 3.3, 3.5, 3.6, and 3.7. Among the verified sources, there is no

equivalent for Lula: none transcribes his statements with date and context and classifies them with the same scheme. This is a limit of the body of sources read and not a property of the politicians, and it should be read that way: a comparative conclusion drawn from this asymmetry would be an artifact of the documentation. Mota et al. classify the statements using Wardle and Derakhshan's typology (Section 3.3), in which disinformation is distinguished from misinformation by the intent to cause harm. Classification, in this scheme, depends on an inference about intent that the record of a statement does not contain; the analysis form replaces it with a question about controlling variables.

Table 3.3 brings together what Lupa's report publishes about the language patterns in each politician's fact-checked statements, with the same taxonomy and the same format.

TABLE 3.3

Language patterns in statements fact-checked by Lupa during campaigns: Lula and Bolsonaro

Pattern in the fact-checked statement	Lula: statements (false; %)	Bolsonaro: statements (false; %)
Strong negations	4 (3; 75.0%)	25 (15; 60.0%)
Absolute terms	7 (4; 57.1%)	39 (21; 53.8%)
Attacks on opponents or parties, crime, or corruption	12 (7; 58.3%)	52 (21; 40.4%)
Electoral process, voting, and voting machines	3 (1; 33.3%)	13 (6; 46.2%)
Numbers, percentages, or amounts	57 (7; 12.3%)	94 (37; 39.4%)
Total statements fact-checked	83 (22; 26.5%)	236 (102; 43.2%)

Note. Agência Lupa (2026, p. 25), "Taxonomia Lupa" [Lupa taxonomy], values read from the image of the page (clipping in verificacao/evidencias/lupa2026-p25-estilo-lula-bolsonaro.png). The categories are not mutually exclusive: the same statement may contain a number, an absolute term, and an attack on an opponent. Only politicians with more than 30 fact-checked statements are included; the statements are from election campaigns, and Lupa does not break down Lula's campaign cycles.

The chapter records what the table allows and what it does not. It allows us to say that, among the statements Lupa fact-checked, those with a strong negation, an absolute term, or an attack on an opponent received "False" in a higher proportion than those containing numbers, for both politicians, and that, among statements with numbers, the proportion of "False" was 12.3% (7 of 57) for Lula and 39.4% (37 of 94) for Bolsonaro. It does not allow us to say that one of them tends more than the other to assert what is false: the bases are selections made by the agency, of very different sizes, the categories overlap, Lula's cells contain 3, 4, 7, and 12 statements, and Lupa itself says it does not rank

who "lied more," because the number of fact-checks depends on time in the spotlight, on the office held, and on communication style. In these terms, the data are the agency's classification; the interpretation is that the form of the sentence correlates with the classification; the hypothesis is that of Section 3.6.

3.9 Rival interpretations and how to decide among them

For a statement classified as false, the analysis form works with six rival interpretations, which Table 3.4 brings together. "Intent to deceive" is not among them, because intent is not an observable controlling variable; what there is are the consequences of the speaker's behavior, the speaker's reinforcement history, and the properties of the occasion. Skinner's definition of a lie is functional: a response emitted in circumstances that would ordinarily control a response incompatible with it (1957).

TABLE 3.4

Rival interpretations for a statement classified as false and what the form would need to record to separate them

Interpretation	Controlling variable at play	What the form would need to record
Weak control by the event (error, slip)	Insufficient stimulation or a poorly conditioned response; no special consequence	Whether the same person describes the event correctly on other occasions, without a special consequence; whether there is a weakness autoclitic
Intraverbal acquired from others	Prior verbal stimulus (text, speech, message)	The source of the chain of words; whether the statement reproduces an earlier wording; absence of a source autoclitic
Distorted tact	Special consequence from the listener	Whether the form changes with the audience; whether correspondence is greater when the audience does not reinforce
Disguised mand or impure tact	The speaker's deprivation or aversive stimulation; what the listener will do next	What the speaker asks of the listener, explicitly or implicitly, on the same occasion (Chapter 5)
Ritual or identity function	Generalized social reinforcement: the exchange is the reinforcement	Whether the content varies without changing the effect; whether the audience reacts to adherence rather than to content (Guerin, 1994)
Effect of selection by the agency	The statement entered the database because it was checkable and circulating	The agency's database and selection method (Chapter 2)

Note. Table by the author, based on Skinner (1957) and Guerin (1994). The six interpretations are not mutually exclusive: Skinner (1957) notes that a response is usually a function of more than one variable.

There are four kinds of observation that the form would call for and that the records used in this chapter do not provide. The first is the occasion: the question, argument, or immediate context to which the statement responds. The second is the audience: who was present and how they reacted, which is the consequence on which the distorted tact depends. The third is the comparison of the same speaker and the same event before different audiences, or before and after the fact-check; it is the only way to separate distortion, through special consequence, from weak control by the event. The fourth is what happens after the correction: whether the form persists (straight repetition), varies (repetition with variation), or disappears. Chapters 6 and 9 deal with what the sources say about audience reaction and about what happens after correction; Chapters 5 and 7 draw from here the distinction among tact, mand, and intraverbal to analyze hate speech and voting.

DATA	Four statements, two by each politician, were classified as "False" by Lupa; a series of statements by Bolsonaro about the vaccine was transcribed and classified by Mota et al. (2023); two of Lula's claims were counted as repeated by Aos Fatos.
INTERPRETATION	The statements are compatible with more than one operant, and the grammatical form of the tact and absolute quantifiers appear in all four. The classification of falsity reports correspondence with a fact, not the controlling variable.
HYPOTHESIS	Some of the statements are a function of special consequences from the audience (distorted tact), some of intraverbals learned from others, and some of noninformative functions (mand, ritual). Separating them requires recording occasion, audience, comparison, and what happens after correction.

Summary

A political statement is verbal behavior, and what identifies it as a tact, mand, or intraverbal is what controls it, not its form. The tact corresponds to an event, and the intraverbal to other words; the community reinforces presenting the second in the form of the first. The distorted tact is a tact whose control by the event has been displaced by the listener's special consequence, in degrees ranging from exaggeration to invention. Repetition is a measure of strength, and the experimental literature suggests that the audience con-

demns less what it has heard before. Source and certainty autoclitics tell the listener where what is said comes from and with how much strength; absolutes and negations ask the listener to accept a state of affairs, and Lupa shows that sentences of this kind receive "False" more often, for logical reasons, for reasons of selection, and, perhaps, for operant reasons.

The form separates data, interpretation, and hypothesis. Applied to four statements by two politicians, it shows what each source documents and what it does not, and leaves open the question that matters: what maintains the statement. Chapters 5 and 7 draw from this chapter the distinction among tact, mand, and intraverbal; Chapter 6, the distorted tact and the verbal subcommunity; Chapter 9, generalized social reinforcement and the audience; Chapter 12 (Sections 12.1 and 12.4), the idea that correcting the statement is acting on a response whose controlling variables may not be the ones the corrector assumes.

References

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Baum, W. M. (1995). Rules, culture, and fitness. *The Behavior Analyst*, *18*(1), 1–21. https://doi.org/10.1007/bf03392688

Catania, A. C., Matthews, B. A., & Shimoff, E. (1982). Instructed versus shaped human verbal behavior: Interactions with nonverbal responding. *Journal of the Experimental Analysis of Behavior*, *38*(3), 233–248. https://doi.org/10.1901/jeab.1982.38-233

Cerutti, D. T. (1989). Discrimination theory of rule-governed behavior. *Journal of the Experimental Analysis of Behavior*, *51*(2), 259–276. https://doi.org/10.1901/jeab.1989.51-259

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Galizio, M. (1979). Contingency-shaped and rule-governed behavior: Instructional control of human loss avoidance. *Journal of the Experimental Analysis of Behavior*, *31*(1), 53–70. https://doi.org/10.1901/jeab.1979.31-53

Guerin, B. (1992). Behavior analysis and the social construction of knowledge. *American Psychologist*, *47*(11), 1423–1432. https://doi.org/10.1037/0003-066x.47.11.1423

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Hayes, S. C., Brownstein, A. J., Zettle, R. D., Rosenfarb, I., & Korn, Z. (1986). Rule-governed behavior and sensitivity to changing consequences of responding. *Journal of the Experimental Analysis of Behavior*, *45*(3), 237–256. https://doi.org/10.1901/jeab.1986.45-237

Menezes, L. F. (2024, December 23). *No segundo ano de governo, Lula desinformou mais sobre fome e educação* [In his second year in office, Lula spread the most disinformation about hunger and education]. Aos Fatos. https://www.aosfatos.org/noticias/lula-2024-desinformou-fome-educacao/

Mota, A. A. S., Pimentel, S. M., & Oliveira, A. V. M. G. (2023). Desordens informativas: análise de pronunciamentos de Jair Bolsonaro contra a vacinação de covid-19 [Information disorders: An analysis of Jair Bolsonaro's statements against COVID-19 vaccination]. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, *17*(2), 311–331. https://doi.org/10.29397/reciis.v17i2.3513

Ribeiro, A. (2023, December 27). *Em primeiro ano de mandato, Lula repete alegações enganosas sobre fome, economia e governo Bolsonaro* [In his first year in office, Lula repeats misleading claims about hunger, the economy, and the Bolsonaro government]. Aos Fatos. https://www.aosfatos.org/noticias/lula-alegacoes-enganosas-2023/

Ribeiro, A. F. (1989). Correspondence in children's self-report: Tacting and manding aspects. *Journal of the Experimental Analysis of Behavior, 51*, 361–367. https://doi.org/10.1901/jeab.1989.51-361

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

instituto WALDEN4 | ACADEMIA WALDEN4

WALDEN4 RESPONDE

The questions that reach the therapist's office, answered

Questions from people seeking therapy, from families who have received a diagnosis, and from students, answered by practitioners who see clients every day, with what research says and what it does not yet know.

60+ answers published

Free access, no sign-up required

walden4.com.br/responde

CHAPTER

4

Political Attitudes and Prejudice as Equivalence Classes

Many attitudes concern things the person has never experienced. Guerin (1994) gives the example of nuclear power: what a person knows about it comes from newspapers and from other people, not from contact with the event. In politics the same seems to hold for much of what gets evaluated, from parties and institutions to groups and to politicians whom the voter knows only by image and by name. The question that organizes this chapter is how a name, an image, or a symbol comes to have an evaluative function for someone who has had no contact with what it designates, and how that function can change.

An experiment from 2021 shows the phenomenon in miniature. Five university students learned, by trial and error, which abstract figures "went with" each of three photographs of Brazilian presidents (Bolsonaro, Lula, and Fernando Henrique Cardoso). Before and after training, they rated photographs and figures on a scale of opposite adjectives. Three of the five rated the abstract figures as neutral on the pretest. On the posttest, the figures came to receive ratings similar to those of the photograph with which they had been associated (Pereira & Moreira, 2021b).

There are five participants, a pretest–posttest design, and one scale. The chapter treats the result as what it is: a demonstration that the mechanism exists under those conditions, not an estimate of how common it is. What the experiment offers is an explicit model of how an evaluation reaches a stimulus without passing through contact with it. Sections 4.1 to 4.3 build that model; 4.4 brings together the studies with socially meaningful stimuli; 4.5 describes in detail the two works with presidents, the 2021b article and the two experi-

ments of the 2021a book, with the stimuli and screens the authors themselves used; Sections 4.6 to 4.8 deal with relations of opposition, attitude change, and the limits of the evidence.

The chapter's figures follow a color code that holds from the first to the last. Blue marks the photographs of the presidents (set A); orange and yellow, the two series of abstract figures (B and C); purple, the party acronyms (D); green, the words *right*, *center*, and *left* (E). Among the arrows, solid dark blue is the trained relation, with correct and incorrect feedback; dashed dark green, the tested symmetry; dashed purple, transitivity. Photographs, abstract figures, and screens come from the works of Pereira and Moreira (2021a, 2021b); the graphs redraw the numbers in the authors' tables; the diagrams made to teach how to read a graph state, on the figure itself, that the values are invented.

4.1 Attitude as evaluative verbal behavior

In social psychology, the core of the definition is evaluation. Ajzen (2001) notes a broad consensus that an attitude amounts to a summary evaluation of some psychological object, along dimensions such as good–bad and pleasant–unpleasant. Bohner and Dickel (2011, p. 392) define it as "an evaluation of an object of thought" and call an attitude object anything a person may hold in mind, from the mundane to the abstract: things, people, groups, and ideas. Albarracin and Shavitt (2018) say that to study attitudes is to study evaluations. The definition was once broader, with cognitive, affective, and behavioral components, and has gradually narrowed to the evaluative component (Mizael et al., 2016b). The dispute that remains concerns the status of the evaluation: Bohner and Dickel (2011) arrange the definitions along an axis that runs from stable entities, stored in memory, to temporary judgments, constructed on the spot.

Behavior analysis shifts the question: instead of what state the evaluation reveals, what contingencies maintain it. For Guerin (1994), attitudes and beliefs are verbal behavior, and shaping by a verbal community is central to how both are formed and maintained. Mizael et al. (2016b) summarize what changes: for the field, attitudes are not merely related to behaviors, they are behavior, overt or covert, emitted or evoked according to a learning history, and, as verbal behavior, they can take on tact, mand, and other functions, depending on the contingencies of the community. This makes it possible to investigate the formation and change of attitudes from relations with the environment, and not as something caused by instances "inside the person" (p. 125). Cameoka and Moreira (2021) write that, understood as verbal behavior,

regardless of topography, attitudes are behaviors maintained by the generalized social reinforcement of the verbal community to which the individual belongs.

For the purposes of this book, I offer a definition in three layers, each adding a restriction to the previous one. The first is that of social psychology: an attitude is the evaluation of an object, with a value that ranges from positive to negative. The second is behavior-analytic: that evaluation is behavior, selected and maintained by consequences, and not a state that causes behavior. The third specifies the type: it is verbal behavior, maintained by the social reinforcement that the verbal community arranges, and therefore it varies among tact, intraverbal, and mand according to the function it serves for the listener (Chapter 3). The first two layers leave the listener out. The third brings the listener back in, and it is through it that a statement of attitude can be read as Guerin read it, by the effect it has on the listener (Chapter 2, Section 2.8).

CONCEPT IN USE

Attitude (Guerin, 1994; Cameoka and Moreira, 2021)

An attitude is the response of assigning to an object (a thing, person, group, or idea) a value that ranges from positive to negative, when that response is verbal: maintained by the consequences the verbal community arranges for it and therefore able to function as a tact, intraverbal, or mand. The object need not have been contacted: control may lie in other words and images. An attitude differs from a belief because the speaker can sustain it in the face of contrary evidence (Chapter 2, Section 2.8).

Two caveats about usage. The first concerns the word *prejudice*, which here follows Pereira and Moreira (2021b): an attitude in which the evaluative attributes are negative. Mizael and de Rose (2017) describe it as a set of negative attitudes toward individuals with certain physical characteristics, skin color, or ethnicity. The label is the researcher's and applies to a relation (group–negative attribute) that the researcher chose to measure; the label is kept with that restriction, just as "hate speech" is kept as an external label (Chapter 5). The second caveat concerns the measure. What the experiments record is a mark on a scale or a choice on a screen, directed to a researcher. That is verbal behavior under the control of the person asking (Guerin, 1994), and Section 4.8 returns to what can and cannot be concluded from it.

4.2 How a word acquires a function without contact with the referent

How does a name come to have value for someone who has never contacted what it designates? Dixon et al. (2003) answer with an example chosen for its simplicity. A person may hear, on one occasion, that the members of a national group are heavy drinkers and, on another, that heavy drinkers are a little *dim-witted*. No one taught this person that the members of the group are dim-witted, but, according to the authors, the two trained relations would suffice to derive the relation between the group's name and the attribute; an unfounded prejudice can thus be established by modes of discourse concerning the group, without direct teaching. The article is conceptual: the example is an illustration, not data.

To know whether trained relations actually produce untrained relations, a procedure is needed in which the trained ones are controlled and the derived ones are tested. That is what the stimulus equivalence paradigm offers, and it is from it that the behavior-analytic literature came to understand attitude formation as a network of arbitrary relations between classes of stimuli and evaluative attributes (Mizael et al., 2016b). Two concepts suffice; I present them in boxes because the chapter uses them from here on.

CONCEPT IN USE

Conditional discrimination and matching-to-sample (Sidman and Tailby, 1982)

In a *matching-to-sample* (MTS) trial, a sample stimulus and two or more comparison stimuli are presented, and the choice of one of them is followed by a consequence for a correct response; which one is correct depends on the sample (Mizael et al., 2021b). The performance defines a conditional relation between stimuli: if A1, then B1; if A2, then B2 (Sidman & Tailby, 1982). Figure 4.1 shows a trial, with the stimuli from one of the experiments the chapter describes later.

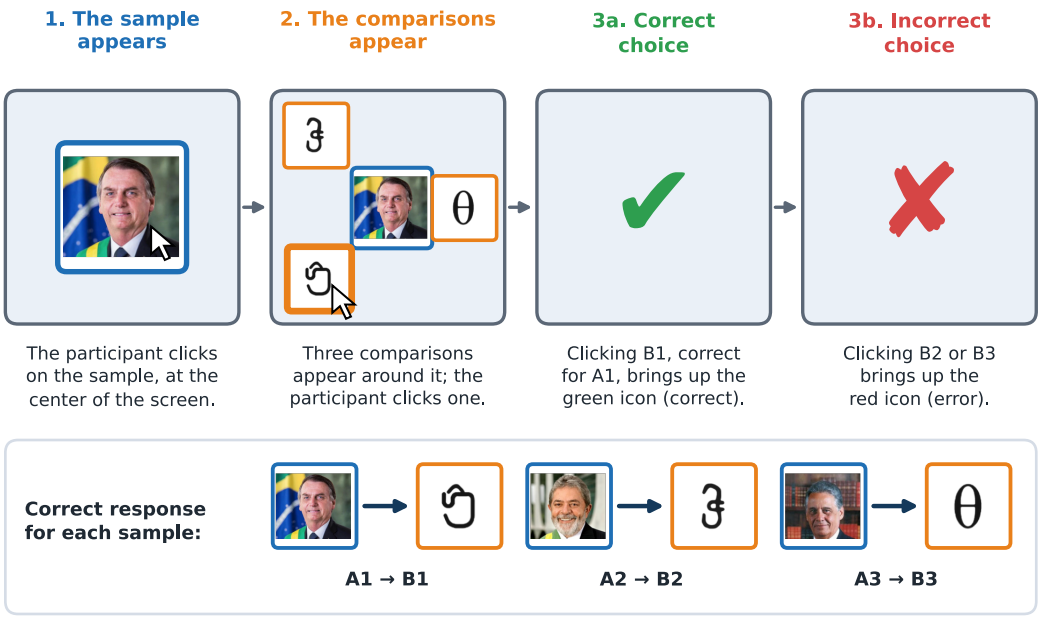
The figure follows a trial of the online experiment described in Section 4.5. In the first panel, the participant sees at the center of the screen the sample stimulus, here the photograph of Bolsonaro, which the authors call A1, and clicks on it. In the second, three abstract figures appear around it, and the participant clicks on one. Which of them is correct is not written anywhere on the screen: it was defined by the researcher, and the participant discovers it only through the consequences. Clicking on B1 in the presence of A1 is followed by a green icon (third panel), and clicking on B2 or B3, by a red icon (fourth panel); after that, the screen goes blank for one second and another trial begins

(Pereira & Moreira, 2021a). The bottom strip summarizes what is taught: in the presence of the photograph of Lula (A2), the correct choice is B2; in the presence of that of Fernando Henrique Cardoso (A3), it is B3. Correct, here, is only what the procedure reinforces: nothing in the task says why that figure would go with that photograph. In the article, the screen was somewhat different (the sample at the top, the comparisons below, side by side, and a happy or sad emoji as the consequence), but the logic of the trial is the same (Pereira & Moreira, 2021b).

Conditionality, however, does not say that the stimuli have any relation other than the one the procedure taught. Sidman and Tailby (1982) observe that equivalence, unlike conditionality, cannot be defined solely by the subject's interaction with the procedure that established it: it requires tests, derived from the three properties that elementary mathematics specifies as the definition of the equivalence relation.

FIGURE 4.1

A matching-to-sample trial, step by step



Note. Assembled with the photographs and abstract figures from the two works of Pereira and Moreira (2021a; 2021b) and with the screen layout of the online experiment (2021a), reproduced in Figure 4.10. "A1→B1" reads as: with A1 as the sample, choose B1.

CONCEPT IN USE

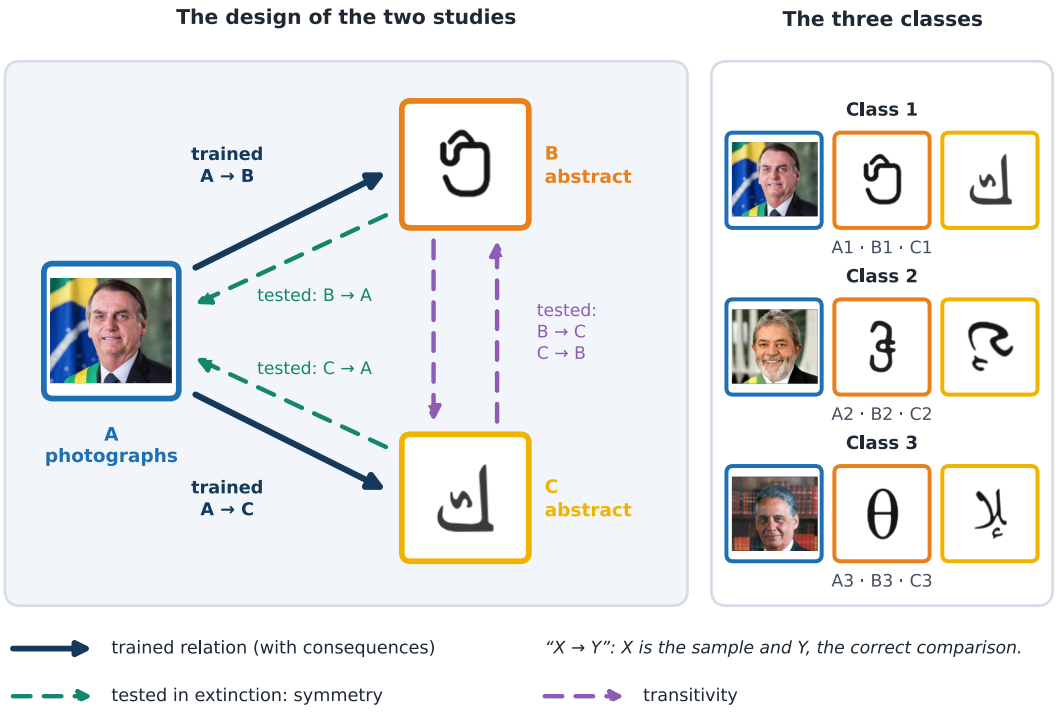
Equivalence class (Sidman and Tailby, 1982)

A set of stimuli forms an equivalence class when the conditional relations among them have the three properties of the equivalence relation. *Reflexivity*: each stimulus is related to itself (A→A), which is tested with generalized identity matching. *Symmetry*: once A→B is trained, B→A emerges without additional training. *Transitivity*: once A→B and B→C are trained, A→C emerges without differential reinforcement. The test that combines symmetry and transitivity (C→A, after A→B and B→C) is the equivalence test (Sidman & Tailby, 1982; Mizael et al., 2016b). In the original paradigm, with three sets of three stimuli, six trained and six emergent relations define three three-member classes (Sidman & Tailby, 1982).

The two studies with presidents, described in Section 4.5, use the paradigm with a variation: instead of an A→B→C chain, two relations are trained from the same stimulus, A→B and A→C, where A is the set of photographs of presidents and B and C are sets of abstract forms (Figure 4.2). No relation between B and C is trained. The relations from B to A, from C to A, and between B and C are tested without consequences, and, when the choices follow the expected arrangement, the three classes, one per photograph, are said to have formed. The relations from B to C and from C to B require combining the return from B to A, which is symmetry, with the step from A to C; in the terminology of the box above, they are equivalence tests, although Pereira and Moreira (2021b) call them transitivity relations.

FIGURE 4.2

Trained relations, tested relations, and resulting classes in the studies with presidents



Note. Design of the studies by Pereira and Moreira (2021a, 2021b), with the photographs and abstract figures from the works themselves (2021a; 2021b). In each training trial, the sample is a photograph (A) and the comparisons are the three figures of a single set (B or C). The dashed arrows indicate relations that the procedure does not reinforce: they are tested without consequences. "X→Y" reads: X is the sample and Y, the correct comparison. Reflexivity does not appear in the design: it is usually assumed, rather than verified, in verbal participants (Mizael et al., 2016b).

Two clarifications prevent misreadings of what is called a class. The first is that the class is defined by tests, not by similarity: the members do not have similar physical attributes and, after training, become substitutable for one another (Cameoka & Moreira, 2021), in certain contexts (Pereira & Moreira, 2021b). The second is that substitutable does not mean equivalent on every occasion: we treat the word "cake" and the cake as equivalent and, even so, we do not eat the word (Mizael & de Rose, 2017). The class is a relation under context, and that is how the term is used below.

The interest of the paradigm lies in what it explains. De Rose et al. (1988) state the problem: a novel behavior, which occurs without an explicit reinforcement history and which is not explained by primary stimulus generalization. To study it, they taught adults, preschool children, and adults with intellectual disabilities two things. The first was a simple discrimination: with forms A1

and A2 on the screen at the same time, choosing A1 was reinforced (A1 was the S+) and choosing A2 was not (A2 was the S−). The second was a conditional discrimination by matching-to-sample: with A1 or A2 as the sample and forms B1 and B2 as comparisons, choosing B1 in the presence of A1 and B2 in the presence of A2 was reinforced, so that B1 and B2 were correct equally often. Afterward, B1 and B2 were presented together, without a sample, and no choice in these probes was reinforced. Given the reinforcement history of B1 and B2, nothing favored either of them, and even so the participants almost always chose B1, the form conditionally linked to A1, the S+ of the simple discrimination. The authors suggest that B1 and B2 acquired, in the conditional training, functions similar to those of A1 and A2. This is the first step of what the next section calls transfer of function.

4.3 Transfer of function

What makes the equivalence class relevant to attitude is transfer of function: the function that one stimulus acquires through direct contact also appears in the other members of the class, without direct contact. Dougher et al. (1994) define transfer of function as a stimulus acquiring a function because it belongs to an equivalence class. Dymond and Rehfeldt (2000) speak of transformation of functions when the functions of one stimulus alter those of another in accordance with the derived relation between them, without additional training. The choice of name has a history: most of the research was reported as transfer, and the authors advocate the generic terminology of relational frame theory (p. 239); Sidman, for his part, objected that "transfer" suggests an unnecessary hypothetical process, and Hayes and colleagues, that the term poorly describes what happens in relations other than equivalence (Weinstein et al., 2008). I use *transfer of function* for what occurs through equivalence and *transformation* when the relation is another one (Section 4.6).

The experiments that support this use arbitrary stimuli and functions that are easy to measure. Discriminative functions extend through the conditional relations, as in the experiment by De Rose et al. (1988) described in the previous section. So do consequential functions. In Hayes et al. (1991), nine introductory psychology students learned, by matching-to-sample with arbitrary visual forms, three three-member classes. Then the comparison stimuli acquired the function of conditioned reinforcer or punisher by being paired with verbal feedback in a sorting task. In eight of the nine participants, the trained function transferred through the respective classes without additional training.

CONCEPT IN USE

Transfer (or transformation) of function (Dougher et al., 1994; Dymond and Rehfeldt, 2000)

Given a stimulus that has acquired a function through direct contact (discriminative, consequential, eliciting, or evaluative), the function is tested in the other members of the class without additional training (Dymond & Rehfeldt, 2000). That the function appeared in the others is data; that it "transferred" through the class is interpretation, to be distinguished from explanations by direct contact (Section 4.8).

So do eliciting functions. In Dougher et al. (1994), eight female university students learned, by matching-to-sample, two four-member classes. One member of one class was paired with electric shock, and a member of the second class was presented with no shock; then all the other stimuli were presented, and conditioning was measured by skin conductance, a physiological measure of emotional response. The response appeared in the other members of the class of the shock-paired stimulus in six of the eight participants. In the second experiment, after all members of one class had been conditioned, one of them was presented in extinction, that is, without shock, and the other stimuli of the class stopped eliciting the response; when the extinguished stimulus was reconditioned, the other members regained the eliciting function.

A finding by Hayes et al. (1991) matters for reading studies with socially charged stimuli. In Experiment 4, even novel feedback stimuli were treated as consequences: the very structure of the task establishes feedback as a consequence. The direction of the effect, however, depended on the transfer of specific functions through the classes. In other words, the procedure already brings functions into the laboratory, and they must be separated from those that the class transferred.

The link with attitudes comes from studies in which the transferred function is evaluative. In Bortoloti and de Rose (2009), university students related, by matching-to-sample, faces with emotional expressions to arbitrary figures and formed three classes, each with one emotional expression and three arbitrary figures. Judgments of the figures on a semantic differential, a scale of pairs of opposite adjectives, became similar to the evaluations of the faces, and even more so when training used delayed matching, that is, with an interval between the sample and the comparisons. In a second experiment, figures one node away from the faces were judged similar to them, and figures three nodes away were not. Nodal distance, that is, how many stimuli separate the

evaluated member from the one that has the function, thus enters as a variable: transfer is not uniform within the class. Weinstein et al. (2008) took the idea to stigma. Fifty university students (45 analyzed) learned by matching-to-sample to respond to arbitrary forms as if they were stimuli related to obesity or to thinness. Then, on an implicit association test (IAT), which estimates bias by how quickly a person associates categories with positive or negative words, with forms similar to the trained ones, a small bias effect appeared (d = 0.31), compared with d = 0.96 on the traditional obesity IAT. d is a measure of effect size, the difference between means in standard deviation units; by convention, 0.2 is small and 0.8, large. The authors attribute the difference to duration: about 45 minutes of training against a lifetime of history.

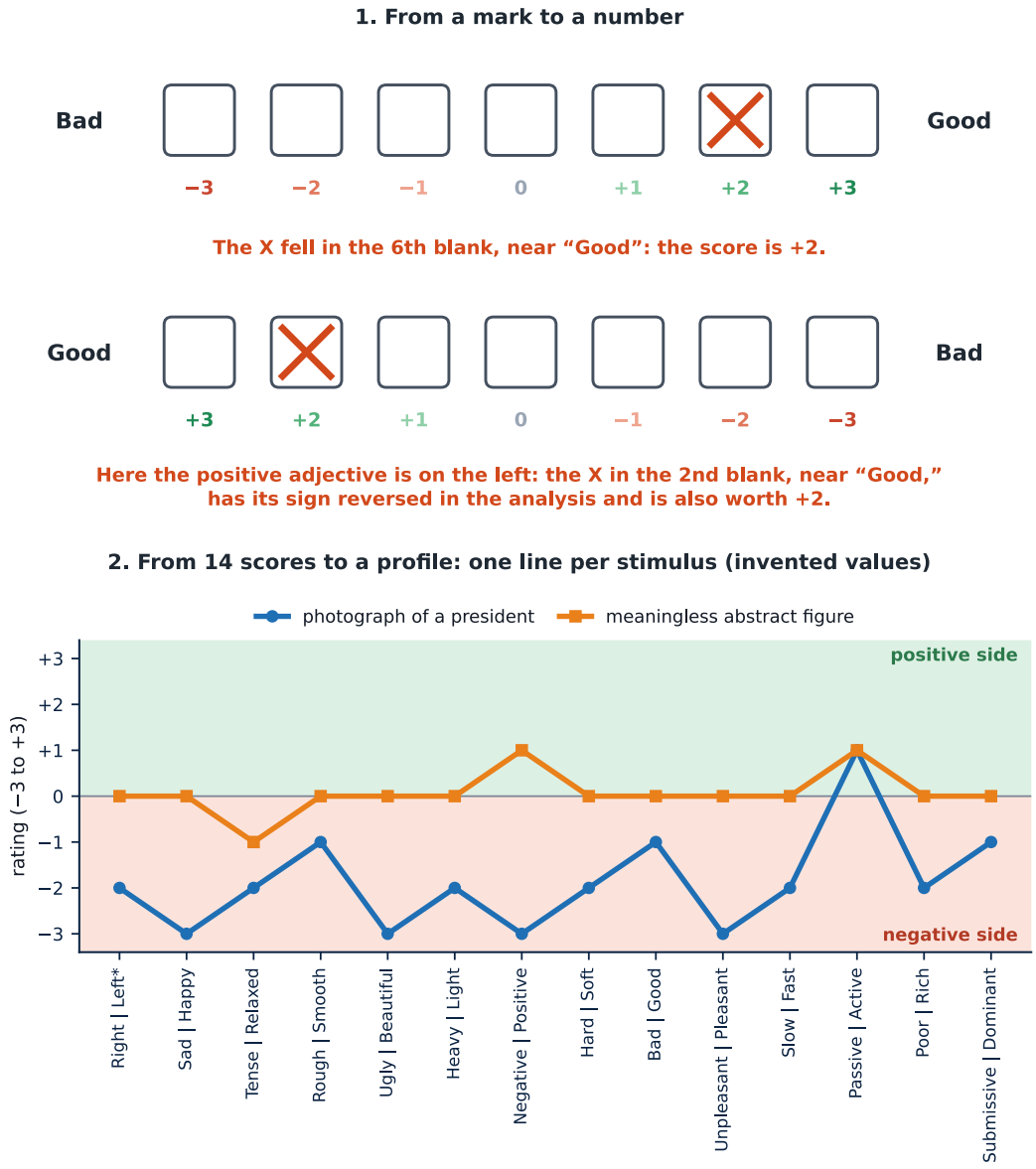
How to read a semantic differential scale

The studies with presidents, which Section 4.5 describes, measure evaluation with a semantic differential scale, and knowing how to read its graphs is what makes it possible to follow the results. The scale of Pereira and Moreira (2021b) had 14 pairs of bipolar adjectives, such as bad and good, with seven blanks between the adjectives of each pair. Each sheet showed one stimulus, and the participant marked, for each pair, an X in the blank that seemed to best describe the stimulus: the one closest to an adjective, if it described the stimulus very well; the middle one, if neither did. In half of the pairs the positive adjective was on the left and in the other half, on the right, so that always marking on the same side does not produce a uniform evaluation.

The top part of Figure 4.3 shows how a mark becomes a number. For the analysis, the authors gave each blank a value from −3 to +3 and reversed the sign of the pairs in which the positive adjective was on the left, so that, in all pairs, the negative adjective was on the −3 side and the positive one, on the +3 side. In the bottom part of the figure, the 14 scores of a stimulus, connected by a line, form that stimulus's evaluation profile. One reads how far the line departs from zero, toward the negative or the positive side, and comparing two lines is comparing two evaluations. The authors chose line graphs because they make it easier to see the "shape of the evaluation" of each stimulus and to compare them, even though the scores of different pairs are not a continuous measure of a single behavior.

FIGURE 4.3

How to read a semantic differential scale: from a mark to a number and from 14 numbers to a profile



** Right/Left: the article converts this pair with its own convention, described in Section 4.8.*

Note. Author's diagram. The 14 pairs are those of the scale of Pereira and Moreira (2021b), with the negative adjective on the left, as in the analysis; for the Submissive–Dominant pair, the article does not say which adjective was treated as negative, and the figure follows the order of the article's list. The values of the two profiles are invented, only to teach how to read the graph. * The Right–Left pair follows a convention of the article's own (Section 4.8).

What the transfer of an evaluation would look like

Figure 4.4 puts the pieces together in a diagram of what transfer of function means for an evaluation. On the left, the three stimuli of a class, taken from the experiment by Pereira and Moreira: the photograph of a president (A1) and two abstract figures (B1 and C1). On the pretest, the photograph has an evaluation of its own, negative in this example, and the abstract figures, which the participant has never seen, stay near zero, the neutral mark, which, as Section 4.5 points out, is not proof of the absence of meaning. In the center is the training, in which A1 is the sample for B1 and for C1, with correct and incorrect feedback, followed by tests without consequences, which check whether the figures lead back to the photograph and whether they relate to each other. On the right, what one expects to see if the function was transferred: the two figures come to have evaluations similar to those of the photograph, which did not change.

FIGURE 4.4

Transfer of an evaluative function through an equivalence class



Invented values, only to show the idea. Each point is an adjective pair (−3 to +3); pink band: negative side; green band: positive side.

Note. Author's diagram, with invented values, only to show the idea; these are not data from any participant. The stimuli are those of the experiment by Pereira and Moreira (2021b), whose training and tests Section 4.5 describes. In the experiments cited in this section with other functions, the event is a shock (Dougher et al., 1994) or verbal feedback (Hayes et al., 1991), and the stimuli are arbitrary forms.

DATA	In experiments with arbitrary stimuli, discriminative, consequential, eliciting, and evaluative functions observed in one member of an equivalence class appeared in other members without direct training (De Rose et al., 1988; Hayes et al., 1991; Dougher et al., 1994; Bortoloti & de Rose, 2009).
INTERPRETATION	Under these conditions, class membership suffices for the function to appear in members that did not acquire it through direct contact. Transfer may vary with procedural parameters and with nodal distance.
HYPOTHESIS	That the same process operates on names, images, and symbols of groups and politicians, in a person's history outside the laboratory, is an extrapolation. Sections 4.4 and 4.5 examine versions of it, with the limits described there.

4.4 Prejudice in the experiments

Prejudice enters equivalence experiments through a difference in material. With arbitrary forms, the participant arrives at the laboratory with no history with the stimuli; with faces, names, emblems, and words that designate groups, that is not the case: the participant arrives with relations already maintained by the verbal community. The studies in this section fall, roughly, into two designs. In one, training teaches relations that do not conflict with that history, and the question is whether the function transfers to arbitrary stimuli (Weinstein et al., 2008; Pereira & Moreira, 2021b). In the other, training teaches relations that contradict it, and the question is whether the class forms and whether the attitude changes. Mizael et al. (2016a) gave this second design the name *Conflicting Relations Paradigm* (Mizael et al., 2016b). Table 4.1 brings together the studies with socially meaningful stimuli whose full text was read; the two works with presidents, which belong to the first design, are described separately, in Section 4.5.

TABLE 4.1

Equivalence experiments with socially meaningful stimuli: what was done, what was observed, and what they do not allow us to say

Study	Participants and stimuli	What was observed	Caveat
Weinstein et al. (2008)	50 university students (45 analyzed); arbitrary forms related, by matching-to-sample, to words about obesity and thinness; measured by IAT	Bias on the IAT with similar forms after training (d = 0.31), compared with d = 0.96 on the traditional IAT	Training of about 45 min; five excluded for low fluency; self-report did not correlate with the traditional IAT
Dixon and Lemke (2007)	12 U.S. university students; photographs of Middle Eastern men, of American men, and of objects; scale from 1 (bad) to 10 (good)	Most reversed the order of their pretest ratings: Middle Eastern men came to receive higher ratings than Americans	Small sample; the bias was probably linked to the 2001 attacks, more intergroup than racial (Cameoka & Moreira, 2021)
McGlinchey and Keenan (1997)	36 participants in Northern Ireland (28 children aged 5 to 11, two 12-year-olds, and six adults); classes with the British flag, an arbitrary stimulus, and an Irish dancer	23 of the 28 children responded in accordance with equivalence (the other five did not on at least one test); the two 12-year-olds and the six adults, on none (13 of 36, or 36%)	Age confounds length of history with other differences (this author's caveat)
Haydu et al. (2015)	28 men from organized supporters' groups of three São Paulo soccer clubs, living in Londrina (Paraná); club emblems, abstract paintings, and the words Good, Bad, and Fair	Baseline and symmetry relations formed; transitivity and equivalence did not emerge; errors, for the most part, in the direction of the history	Only men, from one city; words deliberately assigned to conflict with the history
Moxon et al. (1993)	19 university students (10 men and 9 women); traditionally male occupations, nonsense syllables, and female names; traditionally female occupations as novel stimuli in the test	Equivalence responses on 76% of opportunities among women and 52% among men; novel stimulus chosen on 18% and 42% of occasions; inconsistent generalization, with no difference between groups	N = 19; participants' stereotyping not measured (this author's reading): the explanation by history is the study authors'
de Carvalho and de Rose (2014)	4 children aged 7 to 10 with negative bias against Black men, selected from 16	One reversed the pretest relations; for three, they persisted	N = 4; faces may control responding through expression and other properties (Cameoka & Moreira, 2021)

Study	Participants and stimuli	What was observed	Caveat
Mizael et al. (2016a)	13 children with negative bias against Black faces; forms related to Black faces and to a positive symbol; modified training	All 13 formed the classes; nine maintained the relations on a more stringent test; the difference between ratings of Black and White faces was no longer significant	Four of 13 did not maintain them on the stringent test (this author's calculation)
Mizael et al. (2021b)	46 children with negative bias, in four groups (one control and three with different parameters); six-week follow-up	33 formed classes between Black faces and a positive symbol; 27 maintained at least one relation at six weeks; mean bias positive on the posttest for those who formed classes and negative for those who did not	Association, not cause; the control group had less pronounced bias reduction and maintenance

Note. The numbers come from the articles (Weinstein et al., 2008; McGlinchey & Keenan, 1997; Haydu et al., 2015; Moxon et al., 1993; de Carvalho & de Rose, 2014; Mizael et al., 2016a; Mizael et al., 2021b; Mizael et al., 2016b, for Dixon & Lemke, 2007). The count of 13 of 36 in McGlinchey and Keenan (1997) is this author's and matches the "36% of cases" in the article's text.

Three designs from the table and its surroundings show how these studies are put together. In McGlinchey and Keenan (1997), six adults and thirty children, Catholics and Protestants from Northern Ireland, learned four conditional discriminations: in the presence of A1, a Protestant stimulus (the British flag), choose B1, an arbitrary form; in the presence of A2, a Catholic stimulus (the Irish dancer), choose B2; and then, in the presence of B1 choose C1, a Catholic stimulus, and in the presence of B2, C2, a Protestant one. An equivalence test checked whether the Catholic and Protestant symbols had become related through symmetry and transitivity, and tests with novel stimuli checked whether socially learned relations competed with the emergence of equivalence. Barnes et al. (1996) exposed 12 children with mild intellectual disability and 12 without disability to two equivalence procedures, one with abstract forms, which served as a control, and another with the words "Slow" and "Able" and with personal names, the child's own and a fictitious one, so that training linked the child's name to "Able." The children with disability made fewer equivalence responses on that test, that is, they failed to link their own name to "Able," and the authors suggest that pre-experimental relations between their names and the term "Slow" may have disrupted the formation of the induced relations. In Merwin and Wilson (2005), 32 participants first completed two self-report measures, of distress and of self-esteem, and were divided into high or low distress and self-esteem groups; then they did two matching-to-sample tasks, one that related self-referential stimuli (I, me) to

positive evaluative words and another that related them to negative evaluative words. Those who reported high distress and low self-esteem made more errors on the equivalence tests of the task with positive words.

The table does not show one result; it shows a pattern with three parts. First: with arbitrary stimuli, evaluative functions appear after brief training, with small effect sizes (Weinstein et al., 2008; Bortoloti & de Rose, 2009). Second: when the trained class contradicts social history, formation is frequently disrupted, and the disruption tracks the strength of the history: in McGlinchey and Keenan (1997), the individuals with the most extreme social learning histories, contrary to the training, were the least likely to respond in accordance with equivalence. In Moxon et al. (1993), with occupations and female names, men chose the novel stimulus more than women did (42% and 18% of occasions), and the authors attribute the difference to greater control, among men, by socially learned stereotypical classes; the study did not measure that history. Barnes et al. (1996) and Merwin and Wilson (2005), described above, point in the same direction, with pre-experimental relations between one's own name and the term *slow* and with reported distress and low self-esteem. Third: the same social content produces opposite results depending on the procedure. From 1 in 4 children with negative bias against Black men who reversed the pretest relations (de Carvalho & de Rose, 2014) to 13 in 13 who formed the classes (Mizael et al., 2016a), with the same type of stimulus; Mizael et al. (2016b) attribute the difference, at least in part, to training parameters, such as moving from the simple to the complex protocol and gradually reducing the programmed consequences in the last phase. The difficulty, therefore, lies not only in the history: it lies in the design as well.

Table 4.1 does not exhaust what has been done in Brazil with this design. Instituto Walden4 brought together, in an open-access book, experiments of the same type with socially meaningful stimuli: *Preconceito: doze experimentos e um paradigma* [Prejudice: Twelve Experiments and a Paradigm] (Moreira et al., 2022b). After a first experiment on stimulus control and attitude measurement, which Section 4.8 takes up again, the volume presents experiments on transfer of function and equivalence with stimuli related to women's clothing, to gender and sports, to Candomblé, to tattoos on medical professionals, to obesity, to the social reintegration of prisoners, to older people in the labor market, to drug users, and to body types (the volume's table of contents). They are left out of the table because this chapter did not read them in detail; for readers who want what the table does not show, they are the nearest starting point.

DATA	With socially meaningful stimuli, the intended class failed to form in some of the participants (McGlinchey & Keenan, 1997: 13 of 36; Haydu et al., 2015: transitivity and equivalence did not emerge; de Carvalho & de Rose, 2014: 3 of 4 did not reverse the pretest relations) and formed in others, with modified training (Mizael et al., 2016a: 13 of 13; Mizael et al., 2021b: 33 of 46).
INTERPRETATION	The formation of classes involving socially charged stimuli depends on the relation between what is taught and the participant's history, and on training parameters. Where the trained class contradicts the history, it competes with relations already maintained by the verbal community.
HYPOTHESIS	Competition grows with the length and strength of the reinforcement history, which would explain the difference between young children and adults in McGlinchey and Keenan (1997). The studies compare ages (McGlinchey & Keenan, 1997) or sexes (Moxon et al., 1993); none manipulates the length of the history.

4.5 Attitudes toward Brazilian presidents

Two works by Pereira and Moreira (2021a, 2021b) are, among the equivalence studies checked, the only ones with partisan political stimuli; the authors say they know of no other (2021b). One is the article in *Acta Comportamentalia* (2021b), with five psychology students, a paper scale, and training on a classroom computer. The other is the book (2021a), with two experiments of five students each, run from home, over the internet. This section describes both with the detail of a method report, one at a time, in the order in which the participant experienced them: what they saw, what they did, and what the numbers show. Figures 4.5 to 4.13 accompany the text.

The article: question, participants, and design

The article's question was whether a matching-to-sample procedure could induce the transfer of function from stimuli with political meaning, the photographs of the presidents, to stimuli without meaning, the abstract figures. The authors expected that an affirmative answer would support the equivalence paradigm for the study of attitudes and add evidence with socially meaningful stimuli not linked to prejudice. The independent variable was conditional discrimination training; the dependent variables were class formation, measured by the percentage of correct responses on the tests, and transfer of func-

tion, measured by the difference between the ratings of the abstract figures in the two administrations of the scale. The design is single-subject, with a pretest and a posttest: each participant serves as their own comparison.

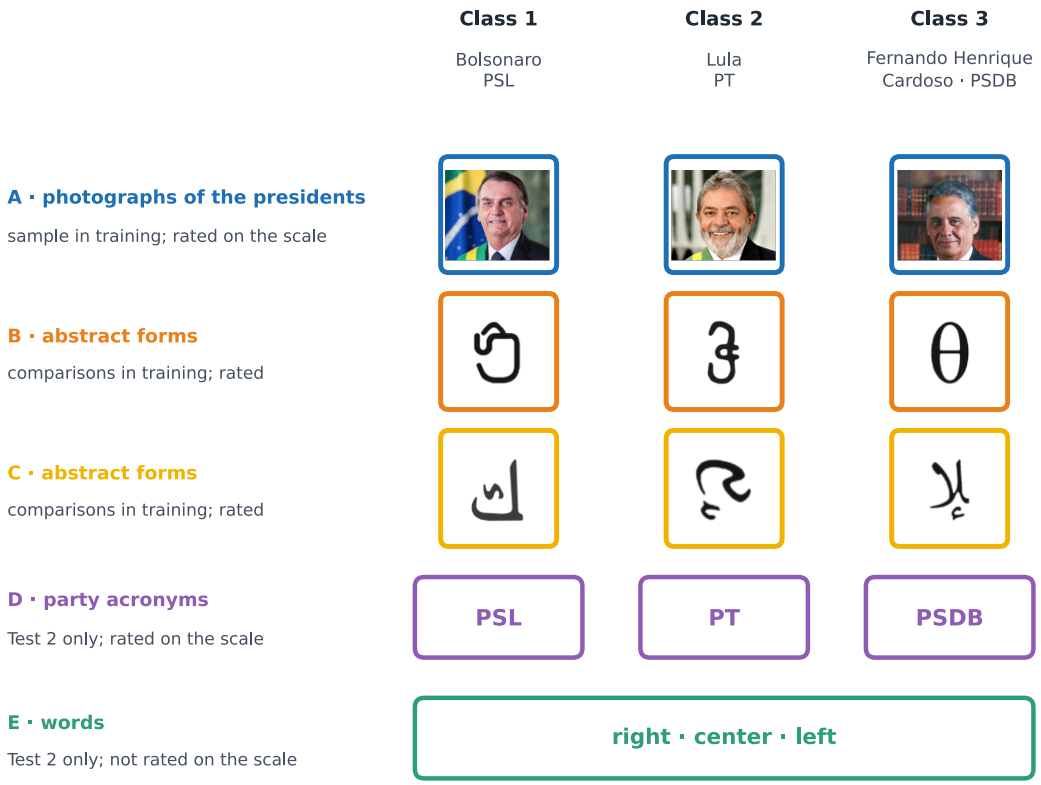
The participants were five university students aged 20 to 25, from the fifth to the seventh semester of a psychology program, chosen by convenience, with no prior contact with the task; all signed the informed consent form, and the project was approved by a research ethics committee. Data were collected in a classroom at a higher education institution, with a computer connected to the internet, a desk, and a chair; the matching task ran on online software developed by the second author, which presented the stimuli and recorded the responses. Each session lasted 80 to 100 minutes.

The stimuli and the trained relations

The stimuli form five sets (Figure 4.5). Set A comprises the photographs of the three presidents: Bolsonaro (A1), Lula (A2), and Fernando Henrique Cardoso (A3). Sets B and C are two series of three abstract figures, with no known meaning for the participants; D contains the acronyms of the presidents' parties, PSL, PT, and PSDB; and E, the words *right*, *center*, and *left*. What the participant learns is to relate A to B and A to C. Each photograph thus has an intended class, formed by it and by one figure from each series: A1, B1, and C1; A2, B2, and C2; A3, B3, and C3. The acronyms and words were not trained: they enter only in the second test, to check whether the participant already related them to the photographs and figures. Only sets A, B, C, and D were rated on the scale, for a total of 12 stimuli.

FIGURE 4.5

The article's stimuli, by set and by class



Note. Photographs and abstract figures reproduced from Pereira and Moreira (2021a; 2021b). The acronyms and words were typeset for the figure. The article does not say which word corresponds to which column (E1, E2, E3), and for that reason set E spans the three columns.

The seven stages

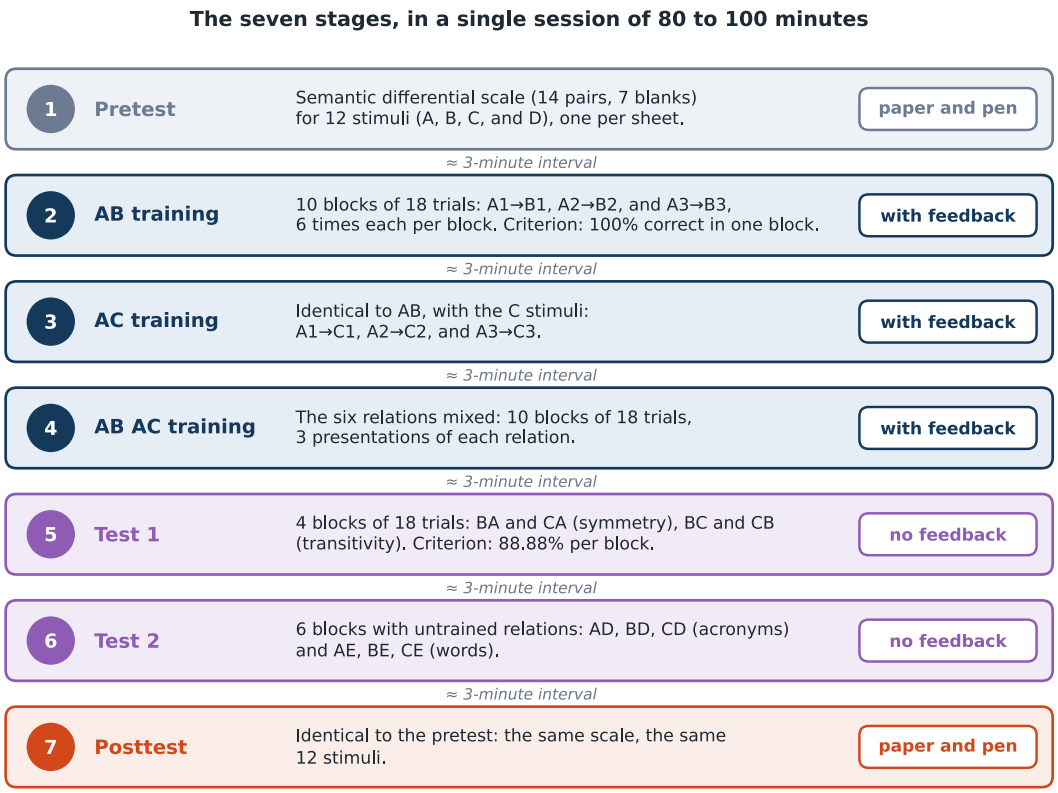
The procedure had seven stages, in a single session, with intervals of about three minutes between them, the time it took to program each stage on the computer (Figure 4.6). On the pretest, the researcher explained orally how to fill out the scale and handed the participant a pad of 12 sheets of A4 paper, one for each stimulus, rated only once, in a fixed order that began with A1, C3, D2, and B1; on each sheet, the participant marked an X on the seven-point scale for all 14 pairs of adjectives.

The three training phases used simultaneous matching-to-sample. In each trial, the sample, a photograph, appeared at the top center of the screen; clicking on it made three comparisons appear at the bottom, side by side, and the sample remained on the screen. After a click on a comparison, all stimuli left the screen and, for one second, a happy emoji appeared if the choice was the one programmed as correct, or a sad one if it was not, followed by two seconds of blank screen. AB training had 10 blocks of 18 trials, with each of

the three relations (A1 with B1, A2 with B2, and A3 with B3) presented six times per block, in random order; the criterion was 100% correct in one block, reached within 30 minutes, and anyone who did not reach it would redo the training on another day. AC training repeated AB training with the C figures, and AB AC training mixed the six relations, in 10 blocks of 18 trials, with three presentations of each.

FIGURE 4.6

The seven stages of the article's procedure



Note. Author's diagram, based on the method of Pereira and Moreira (2021b). The whole session lasted 80 to 100 minutes.

During the tests, no one told the participant whether a choice was correct or incorrect: the first test of relation emergence had no differential consequences. It had four blocks of 18 trials, with six presentations of each tested relation, and checked, in this order, the symmetry relations BA and CA (the figure as sample and the photograph as comparison) and the transitivity relations BC and CB (a figure as sample and the figures of the other set as comparisons). The criterion was 88.88% correct in each block, that is, 16 correct out of 18 trials, to state that the relation had emerged. The second test, of six blocks, checked untrained relations with the acronyms and words: AD, BD, and CD, on one hand, and AE, BE, and CE, on the other. The authors caution that the

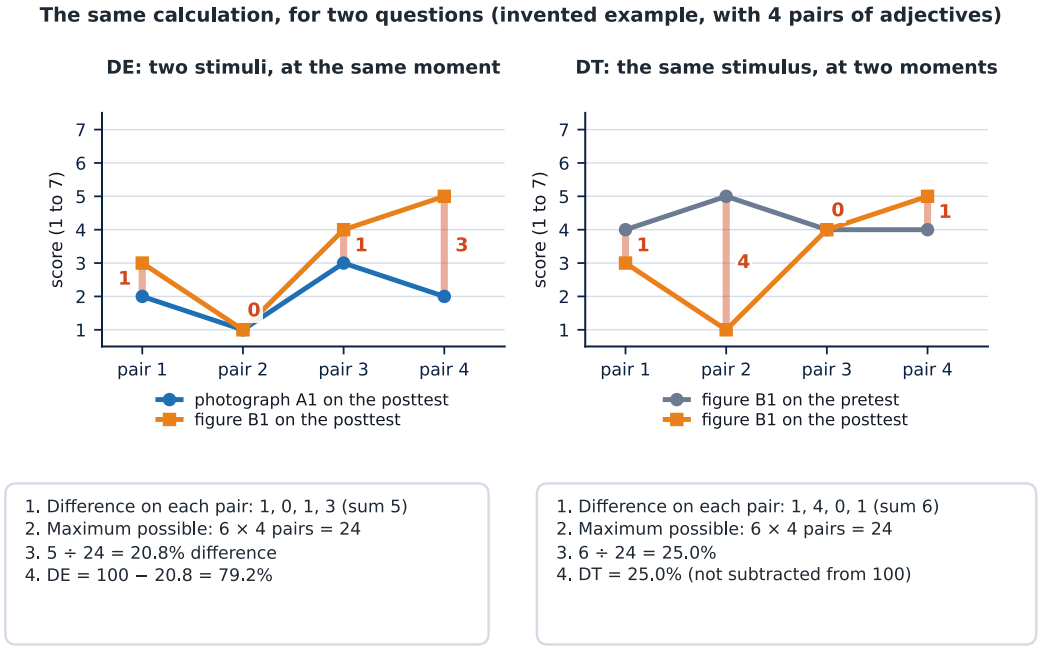
term emergence does not apply to the relations of the photographs with the acronyms and with the words (AD and AE), which probably already existed before the experiment. The posttest was identical to the pretest.

How the Degree of Equivalence and the Degree of Transfer are calculated

To go beyond visual inspection of the graphs, the authors created two percentage measures. The Degree of Equivalence (DE) measures the overlap between the ratings of two different stimuli of the same class on the same test, for example A1 and B1 on the posttest; a DE of 100% means identical ratings on all 14 pairs. The Degree of Transfer (DT) measures the difference between the ratings of the same stimulus on different tests, for example B1 on the pretest and B1 on the posttest; a DT of 100% means that the rating changed completely, for example from −3 to +3 on all 14 pairs. The calculation is the same for both (Figure 4.7). The scale is converted from −3 to +3 into 1 to 7; the absolute value of the difference between the ratings of the two stimuli on each pair is taken; the differences for the 14 pairs are summed; since the largest possible difference per pair is 6, the largest sum is 6 × 14 = 84; the percentage of that sum relative to 84 is the DT, and the DE is 100 minus that percentage.

FIGURE 4.7

How the Degree of Equivalence (DE) and the Degree of Transfer (DT) are calculated



Note. Author's example, with four pairs of adjectives and invented values, to show the calculation; the study used 14 pairs, and the maximum sum was 84 (Pereira & Moreira, 2021b).

With this, the authors expected that stimuli with pre-experimental meaning, such as the photographs, would have a DT close to 0%, that is, that their ratings would not change; that stimuli without meaning, such as the abstract figures, would have a DT close to 50%; and that these same stimuli would have a DE close to 100% with the photographs of their classes. A DT of 50% is equivalent, on average, to moving three points on each pair of the scale, whose range is six points, which corresponds, for example, to going from the center of the scale to one of the extremes; the calculation is by the author of this book and does not appear in the article.

What the article found

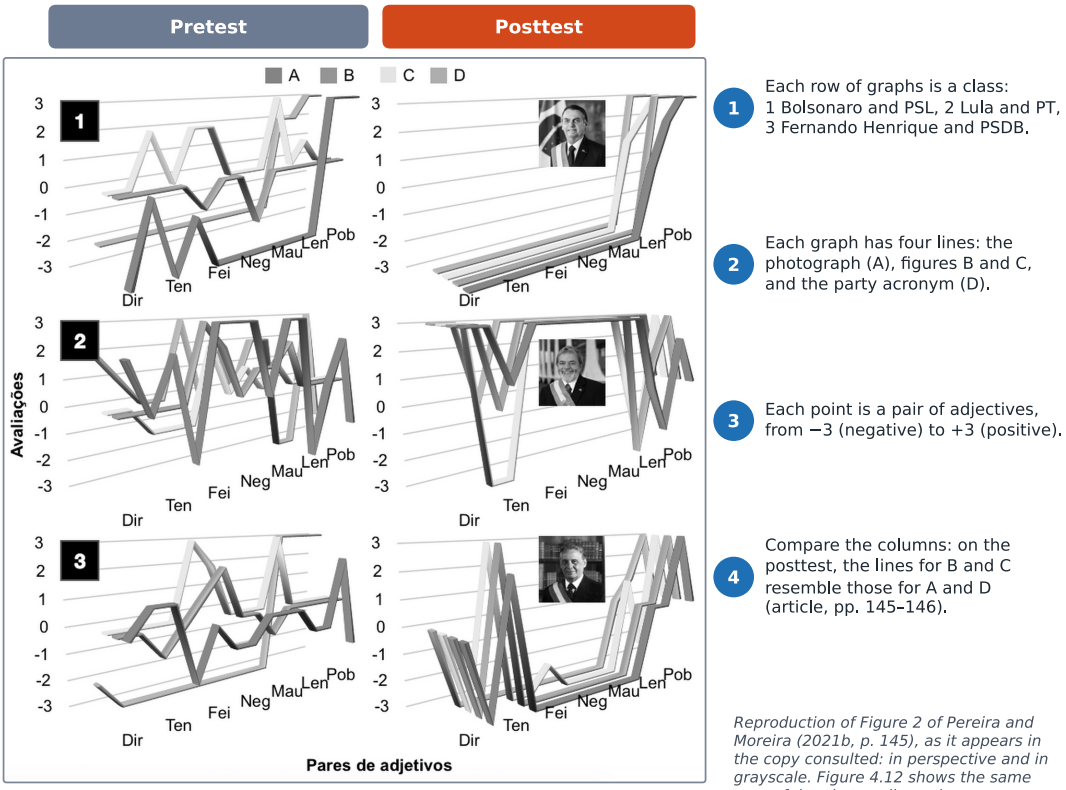
All participants reached the criterion of 100% correct in training (AB, AC, and AB AC), on the first run, and, on the first test, all reached the criterion of 88.88%, except P02 on the BA relation (72.22%); with this caveat, the authors conclude that the three programmed classes formed. On the second test, all scored 94.44% or more. In Table 1 of the article, almost all cells are 100%: the other values are 94.44% (P01 on the CB and BE relations; P05 on CB and BD) and, for P02, 72.22% on BA and 94.44% on CA.

Figure 4.8 reproduces Figure 2 of the article, with P01's graphs. The left column is the pretest and the right column, the posttest; each row of graphs is a class; each graph has four lines, that of the photograph, those of figures B and C, and that of the party acronym. For P01, in class 1, the photograph of Bolsonaro and the acronym PSL received predominantly negative ratings on the pretest, except for the last three pairs; figure B1 stayed near −1 and C1, between zero and three. On the posttest, the photograph and the acronym kept their ratings, and B1 and C1, different from those on the pretest, became similar to those of the photograph. In class 2, Lula and PT were rated predominantly positively on the pretest, and, on the posttest, figures B2 and C2 came to follow the photograph. Class 3 shows another pattern: on the pretest, the photograph of Fernando Henrique Cardoso varied between neutral and negative, and the acronym PSDB was predominantly negative; on the posttest, the acronym kept its rating, and the photograph and the figures became similar to it, which the authors read as transfer from the acronym to the other stimuli of the class.

The results of the other four participants were similar to those of P01 with respect to transfer of function. What varies among them is the starting point of the abstract figures: P02, more than P01, gave them positive and negative ratings already on the pretest, which, strictly speaking, suggests that the figures had pre-experimental meaning for that participant, and P03, P04, and P05 gave them a score of zero, neutral.

FIGURE 4.8

P01's ratings on the pretest and posttest, in the three classes



Reproduction of Figure 2 of Pereira and Moreira (2021b, p. 145), as it appears in the copy consulted: in perspective and in grayscale. Figure 4.12 shows the same type of data in two dimensions.

Horizontal axis: the 14 adjective pairs; vertical axis: the rating, from −3 to +3.

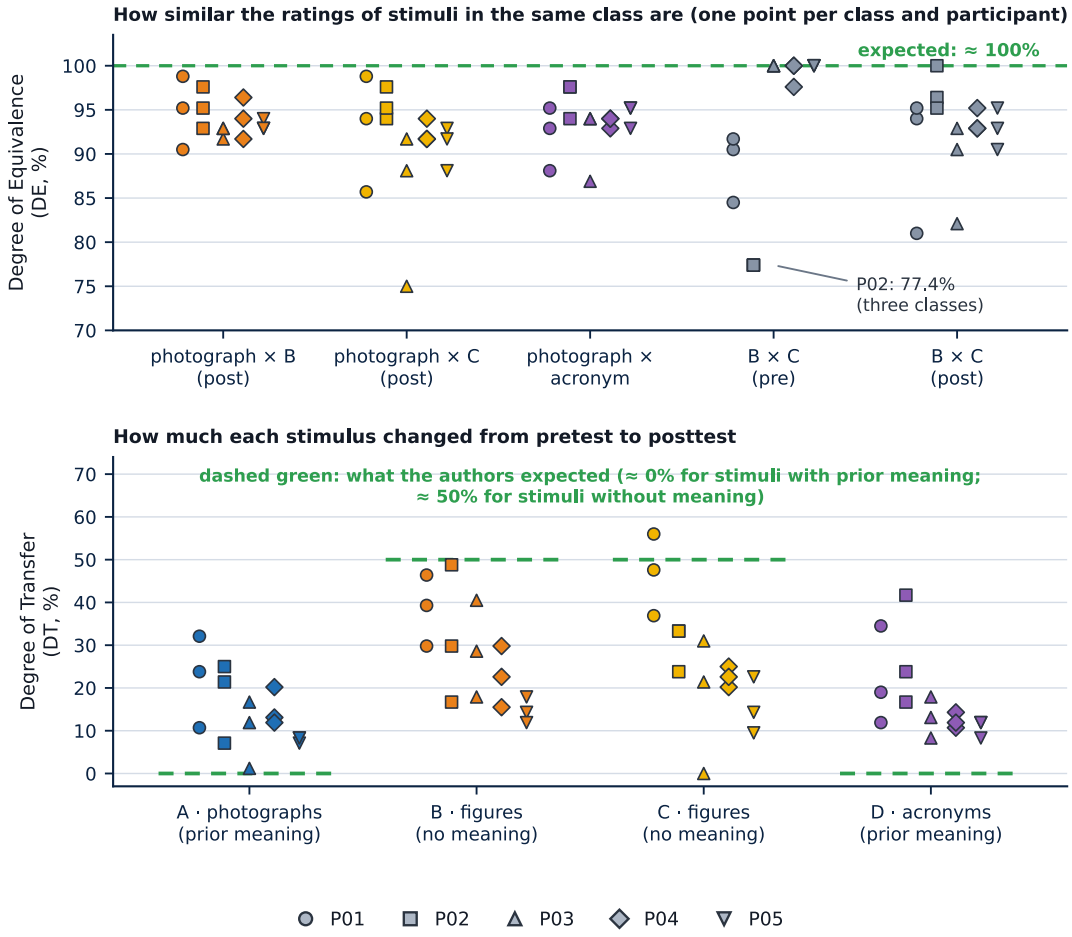
Note. Reproduced from Pereira and Moreira (2021b), as it appears in the copy consulted: in perspective and in grayscale, although the article says the graphs were made in color. The pretest and posttest bands and the reading guide were added. The equivalent figures for the other participants are Figures 3 to 6 of the article. Axis labels are in Portuguese, as in the original: Avaliações (ratings), Pares de adjetivos (adjective pairs); the pair abbreviations come from the Portuguese adjectives.

Figure 4.9 brings together the numbers from Table 2 of the article, for the five participants and the three classes: each column has 15 points. In the top part, the posttest DE between each photograph and the B figure of its class was above 90% for all participants and classes; between photograph and C figure, the pattern was repeated, with differences of less than 5 points relative to that of B, except in P03's class 3 (75.0%, compared with 92.9% for the B figure). The DE between figures B and C, on the pretest, was greater than 90% in most cases, except for P02, for whom it was 77.4% in all three classes. The DE alone could lead to a misleading reading, if the ratings of B on the pretest were already similar to those of A, and for that reason the DTs were also calculated.

In the bottom part of the figure, the DT of the B figures ranged from 11.9% (P05, B3) to 48.8% (P02, B1); that of the C figures, in the table, ranges from 0.0% (P03, class 3) to 56.0% (P01, class 1), and only that value exceeds the 50% the authors expected for stimuli without meaning. The photographs and the acronyms, which were expected to be stable, had DTs of 1.2% to 32.1% and of 8.3% to 41.7%, respectively. For P01, for example, the rating of the photograph of Lula varied by 23.8% and that of Fernando Henrique Cardoso, by 32.1%. The authors admit that it is not clear whether these variations in stimuli with prior meaning are a margin of error of repeated measures or transfer of function from the other stimuli to the photographs, and they speculate that transfer may not be a unidirectional phenomenon.

FIGURE 4.9

Degree of Equivalence and Degree of Transfer for the five participants



Note. Values from Table 2 of Pereira and Moreira (2021b), with no further rounding; each point is one class of one participant. Colors: those of the stimulus sets (Figure 4.5); gray marks the comparisons between B and C. The dashed green lines show what the authors expected.

The book: two online experiments

The book by Pereira and Moreira (2021a) redoes the study in another format. Experiment 1 started from the question of whether it is possible to transfer function from political stimuli to abstract stimuli with a progressive reduction of feedback, proposed in earlier work by other authors: all relations are trained together and feedback is gradually reduced, so as to reinforce intermittently and obtain greater resistance to extinction in the test; the authors say there is only one previous study with political stimuli in this paradigm. The participants were five university students aged 18 to 25, from programs such as psychology, advertising, languages and literature, dentistry, and medicine, invited by WhatsApp message in student groups and chosen by convenience; data collection began only after approval by the ethics committee and the signing of the informed consent form.

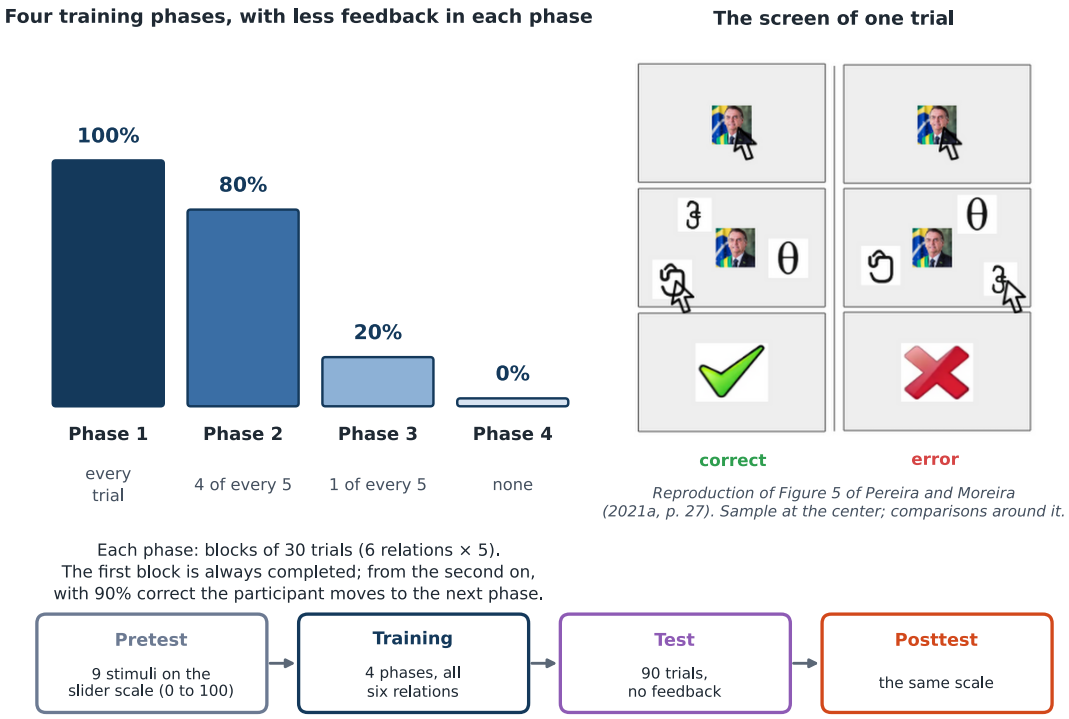
The experiment was entirely online. Each participant was at home, alone in the room, at a computer with an internet connection of at least 10 Mbps; anyone with less would be excluded. A platform presented the trials and feedback automatically and randomly and stored the results. Instructions came in standardized videos, with an animated virtual assistant, in text, and in images; during each activity, the researcher turned off audio and camera, and, between stages, the participant could take a break of up to five minutes. The scale had 13 pairs of bipolar adjectives and was answered on a slider, from 0, closer to the negative adjective, to 100, closer to the positive one, instead of the usual seven points. On the pretest, each participant rated nine stimuli, in random order: three photographs of presidents (A1, Bolsonaro; A2, Lula; A3, Fernando Henrique Cardoso) and six abstract figures (two sets of three); the button to move on was enabled only after 15 seconds and after all pairs had been rated.

Training had four phases, with progressively less feedback (Figure 4.10). In each trial, the sample appeared at the center of the screen, and, after a click on it, three comparisons appeared around it; a correct response was followed by a green icon with a “v” and an error, by a red icon with an “x,” and then by one second of blank screen. In the first phase, feedback came on every trial (100%); in the second, on four of every five (80%); in the third, on one of every five (20%); in the fourth, on none (0%). In each phase, each of the six relations was presented five times, in blocks of 30 trials in random order. The first block was always completed; from the second on, with at least 90% correct the participant moved on to the next phase, and, if the criterion was not reached within 30 minutes, the system ended the task and the participant was disqualified. Finally, the symmetry and transitivity test was a single block of 90 random-

ized trials, without feedback, with five trials per relation: the six baseline relations, the symmetry relations (B and C as sample, A as comparison), and the transitivity relations (BC and CB).

FIGURE 4.10

The book's online experiment: four training phases and the screen of each trial



Note. The screen is a reproduction of Figure 5 of Pereira and Moreira (2021a); the rest is the author's diagram, based on the method of that work.

What the book found

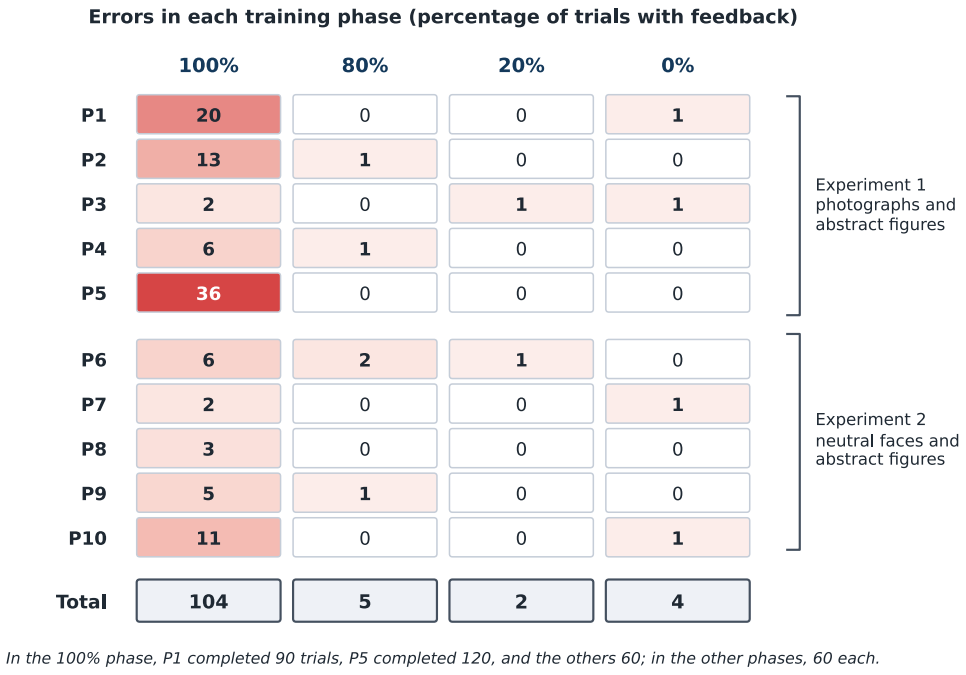
As feedback decreased, almost all training errors occurred when feedback was total. Adding up Tables 1 to 5, from Experiment 1, and 27 to 31, from Experiment 2, there were 115 errors, of which 104 occurred in the 100% feedback phase, five in the 80% phase, two in the 20% phase, and four in the 0% phase (the sum is by the author of this book; Figure 4.11). The largest numbers of errors were those of P5, with 36 errors in the 120 trials completed in the first phase, and of P1, with 20 errors in 90 trials; in the other phases, everyone completed 60 trials and made at most two errors.

On the Experiment 1 test, errors were rare: P1 had one (a transitivity relation, c1b1), P2 none, P3 two (both on b1c1), P4 two (b2c2 and c1b1), and P5 none, so that all participants were correct on more than 90% of the relations in each stage of the test. To find out whether the evaluation was transferred, the authors calculated the similarity between the ratings of two stimuli of a class as

100 minus the absolute value of the difference between the two scores; the closer to 100, the more similar the ratings, and the difference between posttest and pretest similarity, positive if similarity increased, shows what the training did. The relations are called ab (photograph and first set of figures), ac (photograph and second set), and bc (between the two sets of figures).

FIGURE 4.11

Errors of each participant in each training phase



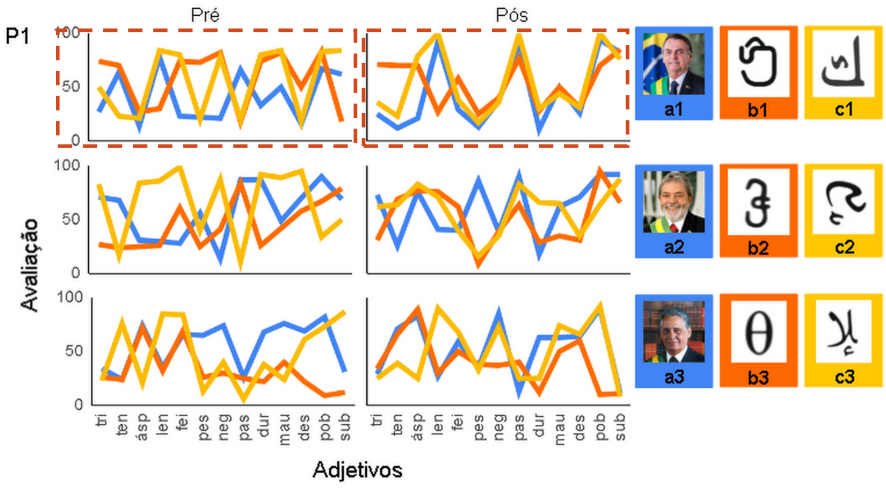
Note. Numbers from Tables 1 to 5 and 27 to 31 of Pereira and Moreira (2021a); the totals per phase are by the author of this book.

Figure 4.12 reproduces Figure 11 of the book, with P1's ratings in the three classes, on the pretest and posttest. The blue lines are the scores for stimulus a (the photograph); the orange ones, those for b; the yellow ones, those for c, the same colors as the sets in this chapter; the 13 pairs of adjectives are on the horizontal axis, and the scores, from 0 to 100, on the vertical axis. The authors report that, for P1, the ratings of the abstract stimuli became more similar to those of the sample in the a1b1c1 relation on the posttest than on the pretest; this is what can be seen in the two dashed boxes, where the three lines diverge on the pretest and run closer together on the posttest. Because the pairs of adjectives had alternating poles, a limitation the authors note, what is compared in the figure is the distance between the lines, not the height of each one.

In Table 26, the mean increase in similarity was 9.4 percentage points for the ab relation, 14.1 for ac, and 6.0 for bc (Figure 4.13). The values per participant vary: the means for the ac relation were between 8.8 and 20.9, those for ab between 4.8 and 12.8, and that for bc was negative for one participant, P5 (−0.9), that is, the similarity between the two series of figures decreased slightly for that participant. The authors also observe that, with the exception of P3, the ratings of the photographs changed less from pretest to posttest than those of the abstract figures, as they expected, because the presidents already had a prior concept; for P3, the similarity of two relations already existed on the pretest, and for that reason they say that, from the graphs alone, it is difficult to assert transfer of function.

FIGURE 4.12

P1's ratings on the pretest and posttest, in the three classes (the book's Experiment 1)



1. Each colored line is a stimulus: blue, the photograph (a); orange, figure b; yellow, figure c.
2. Each point is a pair of adjectives (13 pairs); the height is the score, from 0 to 100.
3. Compare the two dashed boxes (class 1, Bolsonaro): after training, the three lines are closer together (book, p. 48).

Reproduction of Figure 11 of Pereira and Moreira (2021a, p. 48).

Note. Reproduced from Pereira and Moreira (2021a), with the dashed box and the reading guide added. Labels are in Portuguese, as in the original: Pré/Pós (pretest/posttest), Avaliação (rating), Adjetivos (adjectives); the abbreviations come from the Portuguese adjectives.

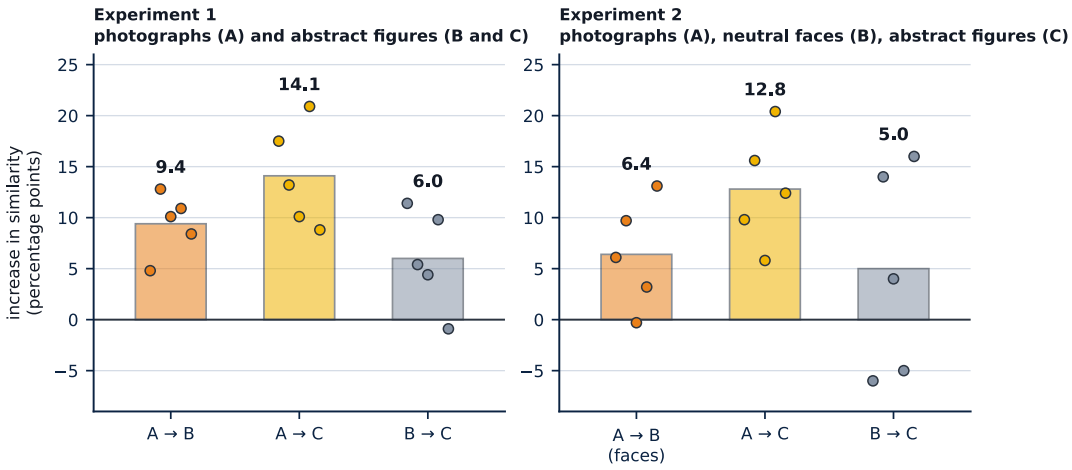
Experiment 2 starts from a hypothesis: the greater the pre-experimental affective load of the stimuli, the greater the difficulty of forming classes and transferring function. The authors illustrate it with a study on jealousy, in which seven of 17 university students did not form classes, and they set out to verify class formation and transfer with faces of ordinary people, rated as neutral, and to investigate whether transfer is easier for abstract stimuli than

for human figures. The five participants had a profile and recruitment identical to those of the first experiment. Set B consisted of faces of unknown White people, chosen because they had been rated as closest to neutral in Cameoka and Moreira (2021), on a 10-point scale (B1 and B2 with a mean of 5.5 and B3 with 5.4); set C consisted of abstract figures, and the procedure was identical to that of Experiment 1.

On the Experiment 2 test, P6 made six errors in 90 trials (one on identity, one on symmetry, and four on transitivity, 93.33% correct overall), P7 and P10 none, P8 one, and P9 one. In Table 52, the mean increase in similarity was 6.4 percentage points for the ab relation, between the photographs and the faces; 12.8 for ac, between the photographs and the abstract figures; and 5.0 for bc. The increase for ac was twice that for ab. By participant, P8 had a mean of −0.3 for ab and 5.8 for ac, and, in that participant's graph, the comparison ratings remained neutral on the posttest as well; for P9, transfer was mainly from the photograph of Lula to the figures of class 2, less so in classes 1 and 3, and was not identifiable for the faces.

FIGURE 4.13

Increase in similarity between ratings, from pretest to posttest, in the two experiments



Note. Values from Tables 26 and 52 of Pereira and Moreira (2021a): bars, the total for the five participants; points, each participant's mean. Colors: those of the stimulus sets, with gray for the relation between B and C.

For the authors, the larger increase for the ac relation than for ab, 50% larger in the first experiment and 100% larger in the second, indicates that it is easier to change the attitude toward abstract symbols than toward neutral faces. They propose that the faces, even when chosen for being neutral, have characteristics, such as being of a man, White, with a beard and hair, that give them a greater pre-experimental load than that of the abstract stimuli, which may

have made transfer more difficult. The limitations they note are online data collection, dependent on internet speed and carried out in home settings, and the pairs of adjectives with alternating poles; each experiment was carried out in a session of one hour to one and a half hours.

What the two works allow us to say

Two findings of the article matter for the chapter's argument. The first is that the party acronyms and the words *right*, *left*, and *center* were correctly related to the abstract figures on the tests, even though no conditional discrimination with them had been programmed; for the authors, this suggests that, before the experiment, they were already part of a class with the photographs of the presidents. On the pretest, the participants systematically rated the photographs of Bolsonaro, Lula, and Fernando Henrique Cardoso as, respectively, right, left, and center. The chapter's reading of this, based on the authors' inference, is that the ideological label is part, for these participants, of the class with which the politician's name already arrives at the experiment. The second finding is a caution from the authors: neutral ratings, around the center of the scale, are not an absence of meaning or of attitude, and P02 gave positive and negative ratings to the abstract figures on the pretest, which suggests pre-experimental meaning for them.

The authors also point out what is missing. They propose replicating the procedure with other training structures, linear series instead of one-to-many, to evaluate the effect of nodal distance, since the expected similarity between B and C holds in part because the distance from A to B and from A to C was the same. And they suggest replicating the work with photographs of anonymous people in place of the abstract figures, which would demonstrate the formation of attitudes toward people with whom the participant has never interacted, a phenomenon directly linked to prejudice, stereotypes, and hate speech. That is, broadly speaking, the design of the book's Experiment 2, with faces in place of figures.

DATA	In five students, after they learned relations between photographs of three presidents and abstract figures, the ratings of the abstract figures on the semantic differential moved closer to those of the corresponding photograph; the Degree of Equivalence between photograph and figure B exceeded 90% for all five participants and all three classes (Pereira & Moreira, 2021b). In the same authors' book, in ten participants in two experiments, the increase in similarity was smaller for neutral faces than for abstract figures (Pereira & Moreira, 2021a).

INTERPRETATION	The evaluation that each participant already made of each politician came to be emitted also in the presence of stimuli with no history with that politician, after a session of about an hour and a half. The design is a single-subject pretest–posttest design, with no untrained control condition; the authors cannot say whether part of the differences in stimuli with prior meaning is variation from repeated measures.
HYPOTHESIS	That transfer is smaller for stimuli with a greater prior affective load, as the authors of that book suggest, is compatible with Haydu et al. (2015) and with Merwin and Wilson (2005), but it was not manipulated: faces and abstract figures differ in several dimensions. That the right–left label is part of the class with which each name reaches the voter is an indication from this study, in five psychology students, not a datum about voters. Chapter 7 (Section 7.6) takes up the idea for candidates' names and numbers.

4.6 "Us" and "them": relations of opposition and comparison

Equivalence is a relation of coordination, but it is not the only one humans derive. Dixon et al. (2003) point out that reliable patterns of relational responding, when they come under contextual control by cues in the environment (such as the words "more than"), are called relational frames, and they cite sameness, opposition, difference, more than and less than, before and after. Weinstein et al. (2008) make the same point: equivalence is not the only possible functional relation in the transformation of functions. The difference matters because, in a relation other than equivalence, the stimuli do not come to share equivalent functions; that is why the authors prefer "transformation" to "transfer," as it allows a broader description.

CONCEPT IN USE

Derived relations of opposition and comparison (Dixon et al., 2003)

A relational frame is a pattern of relational responding under contextual control by cues in the environment (for example, "is the opposite of," "is greater than"). In *opposition*, two stimuli are the opposite of a third: if A is the opposite of Y and A is the opposite of Z, then Y is the same as Z (Dixon et al., 2003; the authors' illustrative example, without data). In *comparison* (more than, less than), the related stimuli do not acquire the same functions: the relation defines how the function is transformed, and not merely whether it is shared (Weinstein et al., 2008).

Equivalence studies with social stimuli offer clues that opposition is at work. In McGlinchey and Keenan (1997), the participants who did not respond in accordance with equivalence were interviewed: they chose novel stimuli because the ones that would form the expected relation were seen as opposites and therefore did not go together. Some argued that the British flag and the Irish dancer had nothing in common and that, therefore, the other comparison "had to be" the match. The authors raise the possibility that these participants may have responded under a frame of opposition, and they report that an earlier study, by Moxon et al. (1993), tried to train classes with stimuli stereotypically seen as opposites, and several subjects did not respond in accordance with equivalence; the description matches the article by Moxon et al. The evidence is modest: a few verbal justifications, in interviews, after the test.

It is a good point from which to read "us" and "them." In Haydu et al. (2015), the words were assigned so as to conflict with the history of each group of supporters: the emblem of the club they supported with *bad* and that of its main rival with *good*; the most frequent errors were, once again, choosing what the history prescribed. That the history of a supporters' group is a history of opposition between "us" and "them" is consistent with the authors' design, since they chose this population because of declared rivalries between the supporters' groups.

DATA	In McGlinchey and Keenan (1997), participants who did not respond in accordance with equivalence explained, in interviews, that the stimuli of the expected relation were seen as opposites.
INTERPRETATION	When the trained class puts side by side stimuli that the verbal community treats as opposites, some participants respond according to the opposition, and not according to the coordination required by the task.
HYPOTHESIS	In political discourse, the "us"–"them" distinction would function as a relation of opposition: attributing an aversive attribute to "them" would transform the function of "us" in the opposite direction, and two groups opposed to the same "them" would come to be treated as equivalent to each other (the rule of the example in Dixon et al., 2003). There are no data, in this chapter, on discourse; the hypothesis needs dated utterances, like those in the analysis form of Chapter 3, and testing with the opposition procedure.

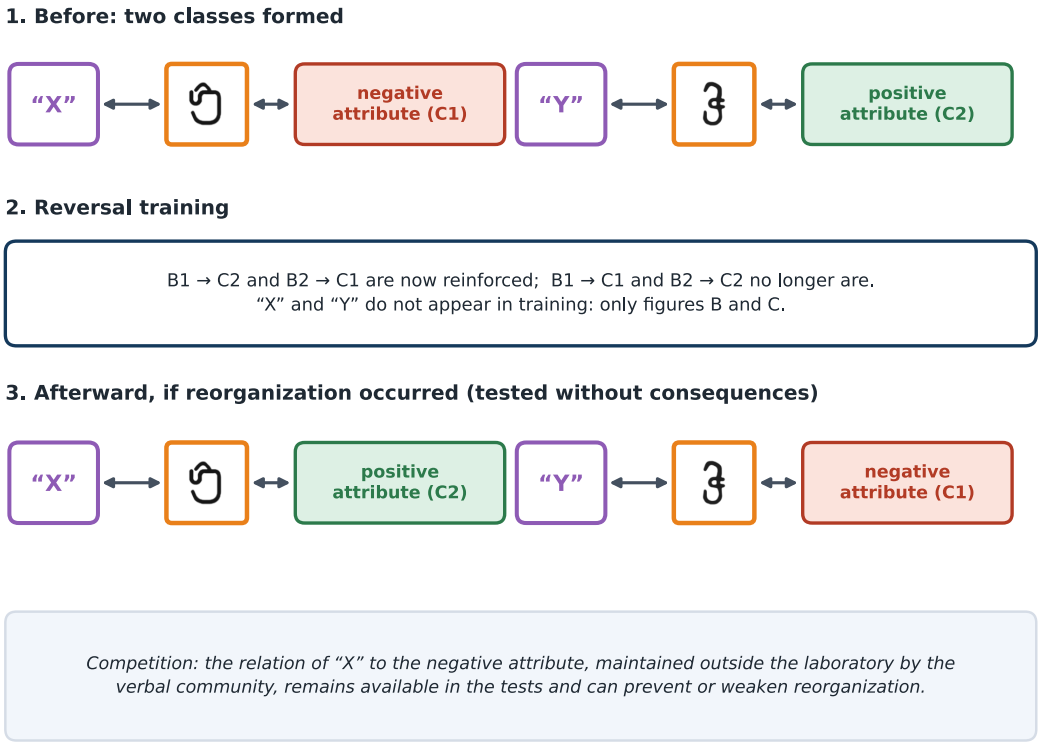
Two observations by Dixon et al. (2003) help anticipate Chapter 12, and they count, like everything in that article, as argument without data. Instructions of the type "don't believe that all members of group X are Y" may heighten the salience of the group as a categorical concept and attach new attributes to the category. And, in the authors' analysis, both prejudice and acts of terror rest on a derived transformation of the functions of whole groups or individuals, grounded in contact, frequently indirect, with a handful of the group's members. Section 5.4 uses this reading for the group label.

4.7 Attitude change: reorganization of classes and comics

In terms of the paradigm, changing an attitude means reorganizing classes. Cameoka and Moreira (2021) describe the reversal of equivalence classes as a type of reorganization in which a rearrangement of the baseline contingencies produces changes in responding consistent with the rearrangement. Figure 4.14 diagrams the procedure with classes formed in the laboratory; in the studies with faces, the class to be undone is the pre-experimental one, and training establishes a conflicting class (Section 4.4).

FIGURE 4.14

Attitude change through reorganization of equivalence classes



Note. The author's original diagram, without data, in the format of reorganization of classes formed in the laboratory (as in Rosendo & de Melo, 2018). In the studies by de Carvalho and de Rose (2014) and Mizael et al. (2016a), AB and BC training creates a class that conflicts with the pre-experimental history.

The evidence is mixed, and the best-supported results are those of partial reorganization. Rosendo and de Melo (2018) studied classes with abstract stimuli, adjectives, and professions associated with the male and female genders. In Stage 1, two experimental groups of eight participants each were exposed to class formation training and tests by delayed matching-to-sample, and class formation was verified in both. In Stage 2, only Experimental Group 1 received reversal training, which rearranges the contingencies learned in Stage 1, and class reorganization tests. A control group of 32 participants underwent no training or testing at all. The three groups rated the professions and the abstract symbols on a semantic differential. In Experimental Group 1, the participants reached the reorganization criterion, of 91% correct, only on the symmetry relations; on the other tests, scores were close to 83% for six participants. Compared with the control group, the ratings on the semantic differential were partially consistent with the classes and with the reversal training, with transfer of function for some scales, and the authors conclude that it is necessary to improve controls over the pre-experimental history with the stimuli and over the number of reversed relations.

With faces and children, the table in Section 4.4 has already shown the variation: one of four children in de Carvalho and de Rose (2014), 13 of 13 with modified training in Mizael et al. (2016a), 33 of 46 in Mizael et al. (2021b). Two points matter for a political reading. One is maintenance: in Mizael et al. (2021b), 27 of the 46 children maintained at least one relation after six weeks, and the control group maintained less. The other is time: Mizael et al. (2016b) argue that training of one or two hours is probably insufficient to modify pre-experimental relations widely reinforced by the verbal community. This is the reviewers' hypothesis, not data from a study that has tested it.

Comics as a change procedure

The studies of Instituto Walden4 replace class training, when the goal is to change an attitude, with a comic strip story. The reason, for the authors of one of the volumes (Moreira et al., 2022a), is resistance: traditional methods of attitude change disregard attitudinal resistance, which narrative-based methods take into account, and narratives, by treating prejudice more subtly, diluted in the plot, are able to minimize the perception of persuasive intent and allow the reader to identify with the characters. The authors found no experi-

ments in the literature on attitude formation after reading comics using the equivalence paradigm and the semantic differential, and that is the gap that the volume's Study 1 sets out to fill.

That study, by Elisa de Pennafort Rabelo, has the design that the subsequent ones repeat. Seven students from any undergraduate program except psychology, from which only first-semester students were included, took part in two sessions, one for each story, no more than three days apart, via Google Meet and Google Forms (Moreira et al., 2022a). Each story, made in Pixton, has seven panels and three female characters: two in a conflict over living together in the house and one in a neutral role; in the first story, Amanda is the only Black character. Before and after reading, the participant rated the three characters of each story (neutral, victim, and aggressor) on a semantic differential scale of 13 pairs of adjectives, from 1 to 7. Reading had an effect on the character ratings of all participants, and most assigned more positive values to the victims, more neutral ones to the neutral characters, and more negative ones to the aggressors, especially after the second story, which may indicate interference from the first session. The authors were unable to check the influence of racial biases and propose, for this purpose, reproducing the experiment with new participants, switching the order of presentation of the stories. Figure 4.17 shows, in the first column, the cast of this study, and Figure 4.18, the conflict of the first story.

Study 2 of the same volume, by Ingrid Taveira Premuli, uses comics on a topic that matters for a political reading: authoritarianism. The stories had characters representing the three factors of the Right-Wing Authoritarianism Scale (authoritarianism, challenge to authority, and traditionalism), and the study investigated, by matching-to-sample, the transfer of function from stimuli with authoritarian meanings to abstract stimuli. According to the abstract, 100% of the sample formed equivalence classes and 80% had their ratings influenced by the stories; the ambivalence in rating, in some participants, leads the authors to question the construction and validity of the semantic differential scale (Moreira et al., 2022a).

The political study in the series is Experiment 8 of Part I (Moreira et al., 2022c), by Luiz Renato Valio Freire, and Figure 4.15 shows it in full. Five adults read, in three segments, an 18-panel story with three characters: C1, with a political ideal leaning to the left; C2, to the right; and C3, with neutral opinions, which serves to observe variables that may be interfering with the research. The characters and panels were made in Storyboard Maker, and the story uses few settings, with few elements, to reduce influences unrelated to what is being studied. In the first segment (Q1), the three characters meet and talk about the weather and the possibility of rain; this is the setting, and the

rating that follows measures the first impression. In the second (Q2), C1 and C2 express opposite opinions on civilian gun ownership and go to opposing demonstrations, one in favor and the other against. In the third (Q3), the three share a ride-hailing car and two flags of political-ideological movements appear in the scene, which, in the figure, read as the Brazilian flag next to C2 and the hammer and sickle flag next to C1. After each segment, the participant rates C1, C2, and C3 on the same 13 pairs of adjectives, from 1 to 7. The comparison among the three tests is the measure of change, with each subject as their own control. Each participant took on average 20 minutes, in a one-week data collection period, and the experiment began only after approval by the Ethics Board.

The method has an ambiguity worth noting. The first stage is called the pretest and described as the presentation of the characters' faces followed by the scale, before the participant learns of their political ideology, but the description of the segments says that Q1 was used before the first rating. Figure 4.15 follows this second description, the more detailed one; in both, the first test comes before any political element, which the description of the stimuli excludes from the first rating.

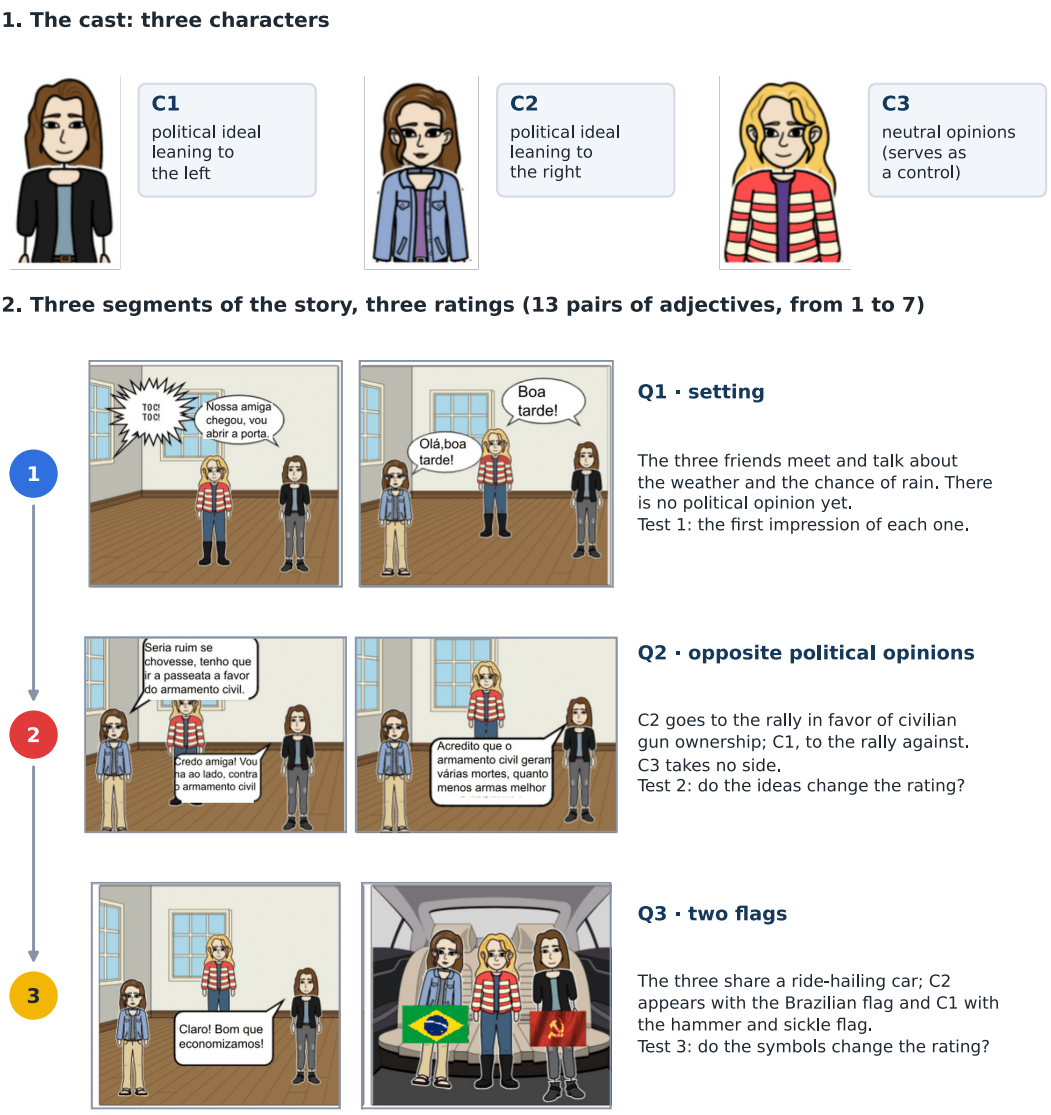
For the graphs, the authors converted the marks from 1 to 7 into values from −3 to +3 and reorganized the pairs that had been presented inverted, so that the negative adjectives were at the bottom; the blue line is Test 1, the red one, Test 2, and the yellow one, Test 3. Figure 4.16 reproduces Participant 1's graph. For C2, the lines for Tests 2 and 3 almost overlap: what changed after the political opinions held after the flags; on most pairs, both lie below zero, and on pleasant–unpleasant they reach −3. For C1 and C3, there were few changes, with exceptions, and, overall, the participant tended to rate C1 more positively and C2 and C3 more negatively across the three ratings.

Among the five participants, there were changes in marks for all characters. C1 received more positive ratings than the other two, and C3, the neutral character, received mostly negative ratings from Participants 2, 3, 4, and 5. Participants 3 and 5 shifted from one pole of the scale to the other, and Participant 2 kept more similar ratings across tests. In spontaneous conversations after the session, which the authors present as anecdotal evidence, the participants said that C3 was rated more negatively because of her passivity and lack of interaction with the other characters, and that the character whose ideology was most similar to the participant's was rated better. For behavior analysis, the first report is instructive: the character designed as a neutral control did not function as a control, because, for the rater, not acting is also something that gets evaluated. The second suggests that similarity between rater and rated may count as much as what the story shows, and a study with five parti-

cipants and no comparison group does not separate one from the other. The authors see the result as agreeing with the group's two previous studies, Rabelo's and Premuli's, that reading comics can produce attitude change toward fictional characters, and they recommend a mixed method, with individual interviews about the reasons for each rating.

FIGURE 4.15

Experiment 8 of Moreira et al. (2022c): three characters, three segments of the story, three ratings

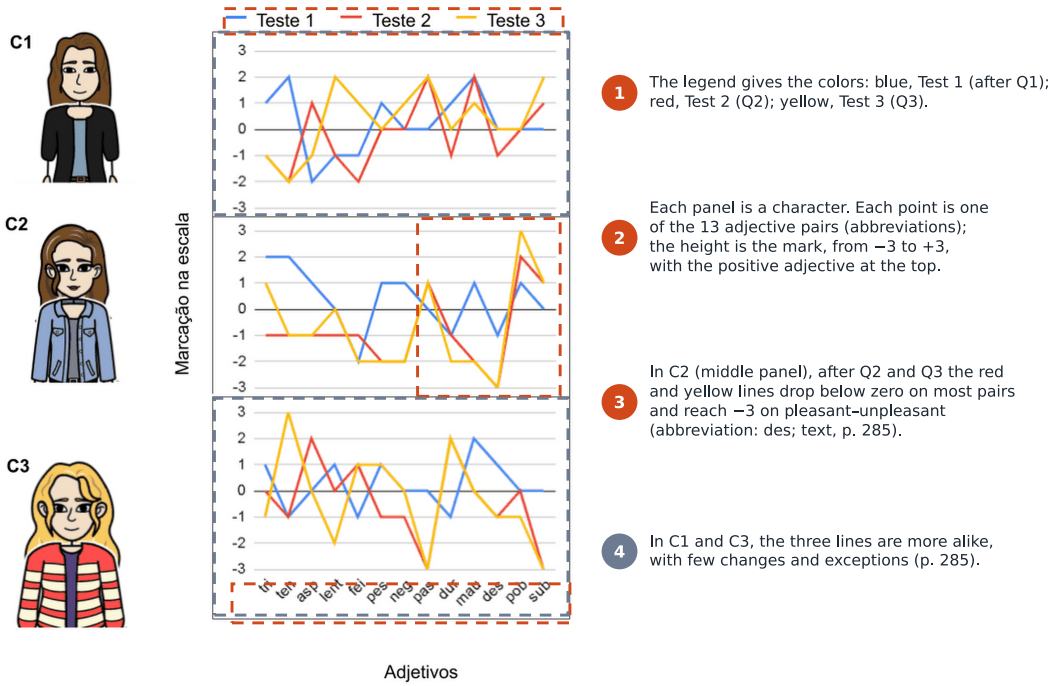


Note. Characters and panels reproduced from Moreira et al. (2022c); the organization into three moments and the side texts are the author's diagram, based on the method. The colors of the tests are those of the lines in the study's graphs (Figure 4.16). The assignment of the flags to the

characters was read from the last image; the study's text says only that there are two flags. The speech balloons are in Portuguese, as in the original study; the side texts summarize each segment.

FIGURE 4.16

Participant 1's ratings in Experiment 8



Reproduction of Figure 4 of Moreira et al. (2022c, p. 286).

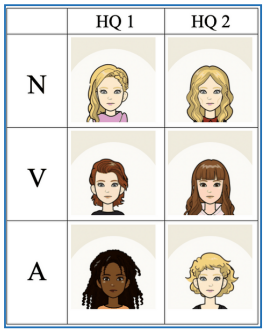
Note. Reproduction of Figure 4 of Moreira et al. (2022c), with the highlights and guide added. The axis abbreviations are those of the study (tri, ten, asp, lent, fei, pes, neg, pas, dur, mau, des, pob, sub, abbreviating the Portuguese adjectives), in the order of the 13 pairs of adjectives, with the negative pole named; the study does not spell them out, and the text reads "des" (from Portuguese desagradável, unpleasant) as pleasant–unpleasant. The legend and axis titles are also in Portuguese: Teste 1–3 (Tests 1–3), Marcação na escala (mark on the scale), Adjetivos (adjectives).

The subsequent volumes repeat the design with other topics and other casts. In Parts IV, V, and VI of the series, each study presents, like Experiment 8, characters, plot, script, the semantic differential form, and tests before and after reading (Moreira et al., 2024c, 2024d, 2024a, tables of contents). Table 4.2 lists the topics, and Figure 4.17 shows the cast of six studies: judging by the labels printed on the figures, it usually includes the one who suffers the act, the one who commits it, and the one who supports, defends, mediates, or observes. Figure 4.18 shows what the participant reads: three scenes, from three books.

FIGURE 4.17

The cast of six studies with comics

Atitudes (2022a), Study 1



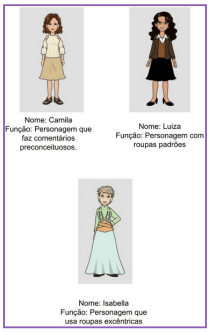
neutral (N), victim (V), and aggressor (A), in both stories · p. 79

Part III (2024b)



psychological violence against women: victim, aggressor, and supporter · p. 92

Part IV (2024c)



clothing: a patient who makes prejudiced comments and two psychologists · p. 111

Part V (2024d)



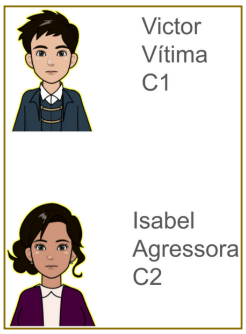
religious intolerance: supporter/mediator, victim, and aggressor · p. 125

Part VI (2024a)



ableism: direct aggressor, covert aggressor, and defender · p. 58

Part VII (2025b)

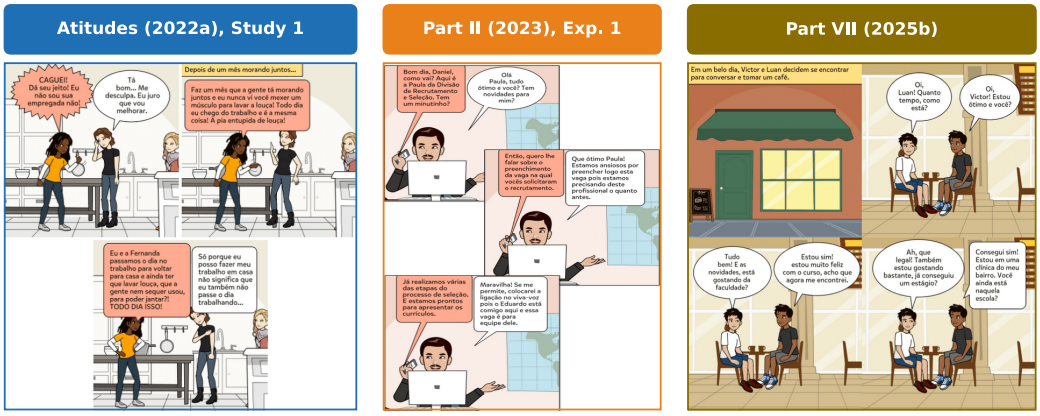


ableism in hiring: victim and aggressor · p. 326

Note. Characters reproduced, with each book's labels, from Moreira et al. (2022a; 2024b; 2024c; 2024d; 2024a; 2025b); the bands and captions are the author's. The topics are those of the study titles (Table 4.2). Labels are in Portuguese, as in each book (Vítima, victim; Agressor/Agressora, aggressor; Apoiador/Apoiadora, supporter; Mediador, mediator; defensor, defender; velado, covert).

FIGURE 4.18

Three comic strip scenes



conflict between roommates, panels 3 to 5 · p. 73

ageism in the labor market: scene 1, panels 1 to 3 · p. 102

ableism in hiring: scene 1, panels 1 to 4 · p. 327

Note. Reproduced from Moreira et al. (2022a), from Experiment 1 of Moreira et al. (2023), and from Moreira (2025b), in that order. What the characters say matters less here than the form: few settings, short lines, and a conflict that the participant reads before rating the characters. The speech balloons are in Portuguese, as in the original books.

TABLE 4.2

Instituto Walden4 books with studies of attitude formation and change, and the topics of the studies

Book	Topics of the studies
Atitudes: histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos [Attitudes: Comics, Semantic Analyses in Social Media, and the Stimulus Equivalence Relations Paradigm] (Moreira et al., 2022a)	Reading comics and fictional characters (Study 1); authoritarianism expressed by characters (Study 2); domestic work (Study 3); verbal behavior on Twitter, the Black Lives Matter movement (Study 4, Chapter 6)
Preconceito: doze experimentos e um paradigma [Prejudice: Twelve Experiments and a Paradigm] (Moreira et al., 2022b)	Stimulus control and attitude measurement (Experiment 1); transfer of function and equivalence with socially meaningful stimuli (Experiments 2 to 12, Section 4.4)

Book	Topics of the studies
Estudos sobre histórias em quadrinhos e atitudes, Parte I [Studies on Comics and Attitudes, Part I] (Moreira et al., 2022c)	Autism at school and in the labor market; transphobia; scientific or pseudoscientific stance; psychologists with religious beliefs; physical attractiveness; former prisoners; political prejudice (Experiment 8); xenophobia; women in video games; plastic surgery and digital influencers; celebrities flaunting wealth
Parte II [Part II] (Moreira et al., 2023)	Validity and reliability of the semantic differential instrument; ageism in the labor market; aggressive, assertive, and passive communication; childhood obesity; clothing; ASD in the labor market; religious prejudice and intolerance; racial discrimination
Parte III [Part III] (Moreira et al., 2024b)	Psychological violence against women in a university setting; pseudosciences; emotional appeal; musical genres; transphobia
Parte IV [Part IV] (Moreira et al., 2024c)	Transphobia; homophobia; discrimination based on clothing; gender in high-performance sports; fatphobia
Parte V [Part V] (Moreira et al., 2024d)	Xenophobia; religiously prejudiced behavior and religious intolerance
Parte VI [Part VI] (Moreira et al., 2024a)	Schizophrenia in the labor market; ableism against autistic people; inclusion of children with ASD; ambivalence in a comic strip narrative
Parte VII [Part VII] (Moreira, 2025b)	Homophobia and ambivalence; racism and homophobia in soccer; gossip news; people experiencing homelessness; mothers in job interviews; alternative treatments for autism; emotional stimuli; hiring people with ASD; social media; inclusive swimming; the Gabriela Pugliesi case; former addicts

Note. Topics as given in the volumes' tables of contents; "Parte II" to "Parte VII" abbreviate the title Histórias em quadrinhos e o estudo das atitudes [Comics and the Study of Attitudes]. Of the studies in the table, only two were read in detail in this chapter, Study 1 of Moreira et al. (2022a) and Experiment 8 of Moreira et al. (2022c).

The volume edited by Moreira (2025a) brings together single-subject studies, mostly with university students, using matching-to-sample and comics, on sexism, homophobia, ageism, transphobia, fatphobia, and prejudice against people with Down syndrome and with autism, in which attitude is measured mainly by a semantic differential scale. The volume's abstract reports that matching-to-sample formed classes and transferred evaluative functions, especially to initially neutral or weakly emotionally charged stimuli, and that there was resistance to changing deeply rooted prejudices or those associated with a strong emotional load (the volume's abstract).

DATA

In reorganization studies with socially meaningful stimuli, reorganization was partial (Rosendo & de Melo, 2018, only for symmetry), varied with training parameters (Mizael et al., 2016a, 2021b), and was partly maintained after six weeks (27 of 46). In the two comics studies read in detail, there were changes in the marks of all participants (Moreira et al., 2022a, Study 1: seven; 2022c, Experiment 8: five), and the character designed as neutral in Experiment 8 received mostly negative ratings from four of the five.

INTERPRETATION	Reorganizing a pre-experimental class is possible under optimized training conditions, but the effect is partial, measured on a scale right after training, and does not say what the person will say in another context. The comics studies examine the effect of information on the rating of fictional characters; with each subject as their own control (Moreira et al., 2022c), the change in marks cannot, by itself, be distinguished from an effect of repeating the measure, and Study 1 of Moreira et al. (2022a) notes possible interference of the first session with the second.
HYPOTHESIS	Attitude change in political discourse would depend, beyond training, on contingencies that continue to reinforce the new relations and stop reinforcing the old ones; the old relations remain reinforced by the verbal community outside the laboratory. This is what Chapter 12 treats as intervention, not as data from this chapter.

4.8 Measures, rival interpretations, and limits

The most frequent attitude measure in the studies of this chapter is a scale: a semantic differential, with pairs of bipolar adjectives, on seven points in the article with presidents and on a slider from 0 to 100 in the book, and on seven points in the studies with comics and with gender and professions; a Likert scale, in Dixon and Lemke (2007) and in Cameoka and Moreira (2021). The authors of the study with presidents take it as a measure of attitude because it relates adjectives to stimuli (Pereira & Moreira, 2021b). There are measures of other kinds: the implicit association test, used by Weinstein et al. (2008), in which self-report did not correlate with the response, and latency-based measures, which Mizael et al. (2021a) include among Brazilian contributions to the assessment of racial bias. None of them ceases to be behavior under the control of the procedure that produces it.

From a behavior-analytic point of view, a mark on the scale is verbal behavior directed to a researcher. Guerin (1994) observes that scales and questions depend on who is asking and on the verbal culture in which the research is done, because attitudes always depend on social consequences, and that the audience functions as a discriminative stimulus for reporting specific attitudes, so that a question can "mand" the answer. Hence the expectation of inconsistencies between what is stated and what is done, especially in verbal behavior. The consequence for the experiments above is direct: a change in the mark is a change in what the person marks, in the presence of that researcher and that scale, and not a demonstrated change in what the person will do or say to other audiences.

The instrument of the two works with presidents has two peculiarities that call for caution. The first is that the numbers from one and the other cannot be compared directly. In the article, the scale had seven blanks per pair, marked on paper and converted into numbers from −3 to +3; in the book, it was a slider from 0 to 100, answered on the computer, with pairs of alternating poles, and the similarity between stimuli was calculated differently, as 100 minus the difference between the scores. The article's DE and DT and the book's similarity measure, broadly speaking, how much two evaluations resemble each other, but the calculations are different. The second is an inconsistency in the article: the method says that, for the Right–Left pair, Right was assigned −3 and Left +3, and the discussion describes the photographs of Bolsonaro and Lula as, respectively, right (values close to 3) and left (values close to −3). The signs in the discussion are the reverse of those in the method. This book records the discrepancy and does not resolve it, because it does not have access to the raw data, and the point the chapter uses, that the participants systematically rated the photographs as right, left, and center, does not depend on the sign.

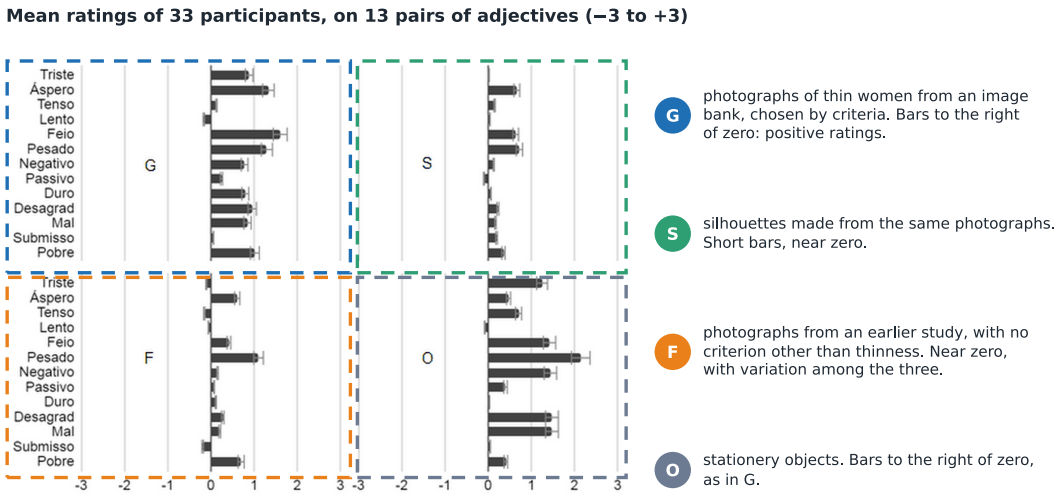
The selection of stimuli also controls the measure. Cameoka and Moreira (2021) conducted three studies, with 75, 40, and 30 university students, in which they rated, on a scale from 1 (bad) to 10 (good), photographs of White men and of Black men. The difference between the means of the two groups of faces was significant in Studies 1 and 2 and not in Study 3, which used the faces of White men rated most favorably and those of Black men rated least favorably in the previous studies; and the mean for the faces of both White and Black men increased significantly from Study 1 to Study 2, in which the faces were slightly smiling. The authors take this as a demonstration of the distorting potential of a loose stimulus selection. The consequence is that "the attitude toward X" measured with a photograph of X is, in part, the attitude toward that photograph.

Experiment 1A of Moreira et al. (2022b) shows the problem in a design built for that purpose. Thirty-three people, aged 15 to 53, rated 12 stimuli in an online form on 13 pairs of adjectives, from 1 to 7: three photographs of thin women from an image bank, chosen by criteria (wearing lingerie or a bikini, facing forward, apparent age of 20 to 30 years, white background and commercial character, and similarity, judged by eight judges, to silhouettes 2 or 3 of the Stunkard Scale); three photographs of thin women from a previous study, with no other criterion; three silhouettes, made from the photographs of the first set, with the body painted black; and three stationery objects. Each stimulus appeared twice, in semirandom order. The photographs chosen by criteria received positive ratings, equal to or greater than 0.8 on most pairs;

the silhouettes, ratings near zero, generally between 0 and 0.4; and the difference was statistically significant on all pairs except Tense/Relaxed and Submissive/Dominant (Figure 4.19). Among the three photographs without criteria, one received more positive ratings, another more neutral ones, and the third more negative ones, and the authors attribute the variation to the only criterion used, that the women be thin, in the researcher's judgment. Because the silhouettes preserve body shape and stayed near zero, the authors suggest that responses to the photographs were not under the control of body shape but of other parts of the stimuli, and they conclude that standardization in stimulus selection is essential for responses to come under the control of the programmed variable, and that skin or hair color, when it is not the variable of interest, can create evaluative biases. For future research, they propose two separate studies, one for stimulus selection and another for the discrimination or equivalence reversal task. For the reading of political utterances, the consequence is the same: the exemplar chosen to represent a group is what gets evaluated, and the attitude "toward the group" is an inference made from it.

FIGURE 4.19

The same concept, four types of stimulus, four evaluation profiles



Reproduction of Figure 3 of Moreira et al. (2022b, p. 47).

Each bar is the mean for one pair of adjectives; the positive adjective is to the right of zero. What changes is the stimulus, not the concept: G, S, and F were meant to represent thin women (pp. 41, 51).

Note. Reproduction of Figure 3 of Moreira et al. (2022b), with the highlights and guide added; each bar is the participants' mean for one pair of adjectives, and the thin lines are those of the original figure (the text does not say what measure they represent). The pair labels are those of the graph, in Portuguese.

What remains is what none of these measures reaches. The experiments deal with individual verbal behavior in the presence of stimuli chosen by the researcher, in short sessions and with small samples, from 4 to 50 participants in the studies of Table 4.1 and five per experiment in those with presidents, often university students or children recruited at schools. Passarelli et al. (2023) distinguish racism at three levels, structural, institutional, and individual, and point to the scarcity of experimental work, partly because of the difficulty of defining it in behavioral terms; what this chapter examined is the individual level. The passage from the laboratory to public discourse is by analogy, and here it is assumed as such: the paradigm offers a model of how evaluative functions reach names without direct contact and of how social history competes with what is taught, and it is with this model that Chapters 5 and 7 read the utterances.

Rival interpretations for the same transfer data

Direct pairing. In the design of Pereira and Moreira (2021b), figures B and C were related directly to the photograph (AB and AC training); the derived relation is BC. Evaluative conditioning, in which direct pairing between a neutral stimulus and a valenced one changes the evaluative response to the neutral one (Mizael et al., 2016b), explains the similarity between A and B, and between A and C, without recourse to the class. The authors propose testing linear series against one-to-many, with nodal distance; Bortoloti and de Rose (2009) have already shown that figures one node from the faces were judged similar and those three nodes away were not.

Control by the audience. The mark may reflect what the participant supposes the researcher expects (Guerin, 1994).

Non-neutral baseline. Neutral ratings are not an absence of meaning (Pereira & Moreira, 2021b), and P02 rated the abstract figures, on the pretest, positively and negatively; the difference between pretest and posttest thus mixes transfer with what the figures already were for that participant.

Repeated-measures variation. The authors admit they do not know how much of the difference in stimuli with prior meaning is variation between administrations of the scale.

Summary

In this book, an attitude is an evaluation of an object that is, at the same time, verbal behavior: maintained by consequences arranged by the verbal community and able to function as a tact, intraverbal, or mand. The stimulus equivalence paradigm offers the model of how an evaluative function reaches a

stimulus without direct contact: once a class forms through conditional discriminations, the function of one member appears in the others, in tests without differential consequences, and varies with nodal distance.

The literature with socially meaningful stimuli shows a pattern. With arbitrary stimuli, evaluative functions appear with small effect sizes after brief training. With stimuli that contradict the participant's history, class formation is disrupted in some cases, and the disruption tracks the strength of the history, which the studies compare (between ages, between sexes) but do not manipulate; with modified training, children formed the intended class and partly maintained it. The experiments with Brazilian presidents show the mechanism in university students, in a pretest–posttest design without a control, with five participants per experiment. In the article's five participants, the ratings of the abstract figures came to resemble those of the photographs (DE above 90%), with moderate changes (DT of 12% to 49% for the B figures); in the book's ten, similarity increased on average by 5 to 14 percentage points, more for abstract figures than for neutral faces. Both works suggest that the right–left label is already part of the class with which each name reaches the participant. In the Instituto Walden4 studies with comics, the marks changed for all participants (seven in one study, five in another), with each subject as their own control and no comparison group; in the political study, the character designed as neutral did not function as a control, and an experiment on stimulus selection showed that the type of photograph, and not only the concept, controls the evaluation.

All these data are marks on a scale or choices on a screen, directed to a researcher, and they admit at least two rival readings: direct pairing between photograph and figure, without a class, and control by the audience. Relations of opposition between "us" and "them" are a promising hypothesis for reading political discourse, supported by only a few interviews and without data on utterances. Chapter 5 uses what this chapter established about transfer for the group label, and Chapter 7, for candidates' names.

References

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52*, 27–58. https://doi.org/10.1146/annurev.psych.52.1.27

Albarracin, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, *69*, 299–327. https://doi.org/10.1146/annurev-psych-122216-011911

Barnes, D., Lawlor, H., Smeets, P. M., & Roche, B. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record*, *46*(1), 87–107. https://doi.org/10.1007/bf03395165

Bohner, G., & Dickel, N. (2011). Attitudes and attitude change. *Annual Review of Psychology*, *62*, 391–417. https://doi.org/10.1146/annurev.psych.121208.131609

Bortoloti, R., & de Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record*, *59*(4), 563–590. https://doi.org/10.1007/bf03395682

Cameoka, M. C., & Moreira, M. B. (2021). Preconceito racial: viés na mensuração de atitudes produzido por controle de estímulos [Racial prejudice: Bias in attitude measurement produced by stimulus control]. *Acta Comportamentalia*, *29*(1), 93–112. https://doi.org/10.32870/ac.v29i1.78782

de Carvalho, M. P., & de Rose, J. C. (2014). Understanding racial attitudes through the stimulus equivalence paradigm. *The Psychological Record*, *64*(3), 527–536. https://doi.org/10.1007/s40732-014-0049-4

de Rose, J. C., McIlvane, W. J., Dube, W. V., Galpin, V. C., & Stoddard, L. T. (1988). Emergent simple discrimination established by indirect relation to differential consequences. *Journal of the Experimental Analysis of Behavior*, *50*(1), 1–20. https://doi.org/10.1901/jeab.1988.50-1

Dixon, M. R., & Lemke, M. (2007). Reducing prejudice towards Middle Eastern persons as terrorists. *European Journal of Behavior Analysis*, *8*(1), 5–12. https://doi.org/10.1080/15021149.2007.11434269

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues*, *12*(2), 129–147. https://doi.org/10.5210/bsi.v12i2.40

Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior*, *62*(3), 331–351. https://doi.org/10.1901/jeab.1994.62-331

Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst*, *23*(2), 239–254. https://doi.org/10.1007/bf03392013

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Haydu, V. B., Camargo, J., & Bayer, H. (2015). Effects of preexperimental history on the formation of stimulus equivalence classes: A study with supporters of Brazilian soccer clubs. *Psychology & Neuroscience*, *8*(3), 385–396. https://doi.org/10.1037/h0101276

Hayes, S. C., Kohlenberg, B., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence relations. *Journal of the Experimental Analysis of Behavior*, *56*(1), 119–137. https://doi.org/10.1901/jeab.1991.56-119

McGlinchey, A., & Keenan, M. (1997). Stimulus equivalence and social categorization in Northern Ireland. *Behavior and Social Issues*, *7*(2), 113–128. https://doi.org/10.5210/bsi.v7i2.310

Merwin, R. M., & Wilson, K. G. (2005). Preliminary findings on the effects of self-referring and evaluative stimuli on stimulus equivalence class formation. *The Psychological Record*, *55*(4), 561–575. https://doi.org/10.1007/bf03395527

Mizael, T. M., & de Rose, J. C. (2017). Análise do comportamento e preconceito racial: possibilidades de interpretação e desafios [Behavior analysis and racial prejudice: Possibilities of interpretation and challenges]. *Acta Comportamentalia*, *25*(3), 365–377. https://doi.org/10.32870/ac.v25i3.61632

Mizael, T. M., Coelho, C. L., Rodrigues, W. C., & de Almeida, J. H. (2021a). Racial issues and behavior analysis: Experiences and contributions from Brazil. *Behavior and Social Issues*, *30*(1), 495–513. https://doi.org/10.1007/s42822-021-00071-1

Mizael, T. M., de Almeida, J. H., Roche, B., & de Rose, J. C. (2021b). Effectiveness of different training and testing parameters on the formation and maintenance of equivalence classes: Investigating prejudiced racial attitudes. *The Psychological Record*, *71*(2), 265–277. https://doi.org/10.1007/s40732-020-00435-w

Mizael, T. M., de Almeida, J. H., Silveira, C. C., & de Rose, J. C. (2016a). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record*, *66*(3), 451–462. https://doi.org/10.1007/s40732-016-0185-0

Mizael, T. M., Dos Santos, S. L., & De Rose, J. C. C. (2016b). Contribuições do paradigma de equivalência de estímulos para o estudo das atitudes [Contributions of the stimulus equivalence paradigm to the study of attitudes]. *Interação em Psicologia*, *20*(2). https://doi.org/10.5380/psi.v20i2.46278

Moreira, M. B. (Ed.). (2025a). *Formação e mudança de atitudes: Pareamento ao modelo e história em quadrinhos* [Attitude formation and change: Matching-to-sample and comics] (1st ed.). Instituto Walden4.

Moreira, M. B. (Ed.). (2025b). *Histórias em quadrinhos e o estudo das atitudes - Parte VII* [Comics and the study of attitudes, Part VII]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_2025_hq7.pdf

Moreira, M. B., Ávila, A., Corrêa, B. F., Reis, F. G. dos, Oliveira, L. I. R. de, Pol, L. V., Silva, L. C. da, & Aguiar, S. H. A. (2023). *Histórias em quadrinhos e o estudo das atitudes - Parte II* [Comics and the study of attitudes, Part II] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2023_hq2.pdf

Moreira, M. B., Cardoso, G. P., Silveira, L. A., Carvalho, L. R., & Rabelo, M. M. (2024a). *Histórias em quadrinhos e o estudo das atitudes - Parte VI* [Comics and the study of attitudes, Part VI]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq6_2024.pdf

Moreira, M. B., Lemos, I. K., Nogueira, L., Silverio, J. P. N. A., Bacelar, J. M. C., Morais, M., Godoy, M. B. C., & da Silva, V. A. L. R. (2024b). *Histórias em quadrinhos e o estudo das atitudes - Parte III* [Comics and the study of attitudes, Part III]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq3_2024.pdf

Moreira, M. B., Martins, A. C. M., Silva, A. F. V., Ribeiro, G. M., Faria, G. M. S., & Bezerra, J. G. N. (2024c). *Histórias em quadrinhos e o estudo das atitudes - Parte IV* [Comics and the study of attitudes, Part IV]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq4_2024.pdf

Moreira, M. B., Povoa, A. M. R., Belarmino, A. J. A., Meireles, A. D. R., & Tanabe, G. Y. R. (2024d). *Histórias em quadrinhos e o estudo das atitudes - Parte V* [Comics and the study of attitudes, Part V]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq5_2024.pdf

Moreira, M. B., Premuli, I. T., Rabelo, E. de P., Pessoa, M. B. F., & Medeiros, M. R. G. C. (2022a). *Atitudes: Histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos* [Attitudes: Comics, semantic analyses in social media, and the stimulus equivalence relations paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_q0.pdf

Moreira, M. B., Rocha, G., Lago, J., Soares, A., Santos, B. dos, Simões, C., Dias, G., Nunes, J., Alves, K., Pires, M., Jardim, M., Chadud, L., Alvarenga, O., & Souza, W. (2022b). *Preconceito: Doze experimentos e um paradigma* [Prejudice: Twelve experiments and a paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_12exp.pdf

Moreira, M. B., Sousa, C. P., Tomasini, C., Craveiro, H. do A., Rodrigues, I. L., Dutra, I., Barros, J. F., Freire, L. R. V., D'arcanchy e Vilela, M. E., Cautela, M. E. T. C., Oliveira, P. M. D. N., Soares, S. K. S., & Andrade, V. M. (2022c). *Estudos sobre histórias em quadrinhos e atitudes - Parte I* [Studies on comics and attitudes, Part I] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_hq1.pdf

Moxon, P. D., Keenan, M., & Hine, L. (1993). Gender-role stereotyping and stimulus equivalence. *The Psychological Record*, *43*, 381–394.

Passarelli, D. A., Rico, A. S., & Silvestre, M. H. (2023). Contribuições da análise do comportamento para a compreensão do racismo [Contributions of behavior analysis to the understanding of racism]. *Revista Brasileira de Análise do Comportamento*, *19*(1). https://doi.org/10.18542/rebac.v19i1.14941

Pereira, P. B. G., & Moreira, M. B. (2021a). *Atitudes em relação a presidentes do Brasil: Formação de classes de equivalência e transferência de função* [Attitudes toward Brazilian presidents: Equivalence class formation and transfer of function] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/pereira_moreira_2021_presidentes.pdf

Pereira, P. B. G., & Moreira, M. B. (2021b). Transferência de atitudes em relação a presidentes da república [Transfer of attitudes toward presidents of the republic]. *Acta Comportamentalia*, *29*(4), 137–154. https://doi.org/10.32870/ac.v29i4.80318

Rosendo, A. P., & de Melo, R. M. (2018). Transferência de função e reorganização de classes de equivalência relacionadas a gênero e profissões [Transfer of function and reorganization of equivalence classes related to gender and professions]. *Revista Brasileira de Análise do Comportamento*, *14*(1), 31–43. https://doi.org/10.18542/rebac.v14i1.7157

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. Matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

Weinstein, J. H., Wilson, K. G., Drake, C. E., & Kellum, K. K. (2008). A relational frame theory contribution to social categorization. *Behavior and Social Issues*, *17*(1), 40–65. https://doi.org/10.5210/bsi.v17i1.406

instituto WALDEN4 | ACADEMIA WALDEN4

EVIDENCE-BASED PRACTICES

What works in autism, practice by practice

One page for each evidence-based practice, drawn from the review by the National Clearinghouse on Autism Evidence and Practice: 31,779 articles screened and 972 analyzed, with the number of studies supporting each one.

28 practices described

Free access, no sign-up required

walden4.com.br/pratica-baseada-em-evidencias/autismo

CHAPTER

5

Hate Speech: Disguised Mands and Aversive Control

In 2011, the TV show CQC, broadcast by the Bandeirantes network, asked Jair Bolsonaro, then a federal deputy, what he would do if he had a gay son. According to Agência Lupa (a Brazilian fact-checking agency), which reproduces in its 2026 report the fact-check it published on August 3, 2018, he answered: "That doesn't even cross my mind, because they had a good upbringing. I am a present father, so I don't run that risk" (Agência Lupa, 2026, p. 24). In November 2017, the Rio de Janeiro State Court of Justice upheld the deputy's conviction for collective moral damages over statements with homophobic content made on the show; the penalty set in 2015 provided for compensation of R$150,000, and the case, at the date of the fact-check, was still pending. "Homophobic content" is Lupa's description, and the conviction is the court's, according to Lupa. Both are recorded here with that attribution, with no classification added.

This chapter starts from cases like this one, that of a statement classified by an external body, and asks what behavior analysis can say about it without repeating the classification as if it were an explanation. Calling a statement hate speech, homophobic, or racist describes what it is for whoever classifies it. It does not say what produces it or what maintains it. The chapter's question is the second one: what contingencies maintain statements of this kind, on the side of the speaker, of the audience, and of the group they refer to. There are three tools: interlocking contingencies (Section 5.1), the disguised mand (5.2), and aversive control (5.3). Then come the group label and transfer of function (5.4), what the data allow us to state about the passage from speech to act (5.5), and the Chapter 3 analysis form applied to the opening case (5.6).

A rule that runs through the whole book applies here with more force than in the other chapters: "hate speech" is an external label. Only a statement classified by an external body, such as a court decision, removal by a platform, or the coding of a scientific study, is called that, with the body named. Without external classification, the function of the statement is described and the statement is not labeled. The definitions and measures each study uses are discussed in Chapter 1; here they are taken as the study presents them.

5.1 What the speaker's behavior produces: approval, removal of threats, and functions that vary with context

To find out what maintains a statement, the analysis asks about the consequences it produces. What the verified sources allow is a short list of candidates, none of them tested with politicians' speech.

The first is the approval or attention of an audience. This is the generalized social reinforcement of Chapter 3 (Section 3.3): for Guerin (1994), attitudes are statements maintained by this kind of reinforcement in a verbal community. Fazzano and Gallo (2015), who propose a functional hypothesis for aggression against homosexuals (Section 5.3), hypothesize, in addition to negative reinforcement, the existence of social reinforcement, provided by peers, through the group's approval. The response class they analyze has diverse topographies (hitting, insulting, torturing, beating, excluding, discriminating) and is defined by function, not by form.

The second is the removal or reduction of aversive stimulation. Skinner (1957) includes escape from and avoidance of aversive stimulation among generalized reinforcers, and among the ways of stimulating someone aversively he lists many sorts of "verbal damage." Fazzano and Gallo (2015) hypothesize that the presence of homosexuals would be aversive to the aggressor and that the aggression would be negatively reinforced, and they observe that a reduction in the intensity of the stimulus is already enough for reinforcement, so that violence need not end in death to be reinforced. That the aversiveness is learned and not a property of the stimulus is their point: there is nothing in the stimulus itself that is aversive, and the case seems to be one of learning. The acquisition processes they point to are shaping by differential reinforcement, instruction following, and imitation of a model, "very possibly" all three together.

The third is that functions vary. Mizael (2018), in a review of the behavior-analytic literature on homosexuality (Section 5.3), when listing what behavior analysis could do for the topic, describes (citing a contextual analysis of racism published by Guerin in 2005) different functions of "being racist" that

would control people's behavior: being accepted, raising the status of one's own group, avoidance, thoughtless use of stereotypes. What the list teaches is that the same statement may have, on different occasions or in different speakers, different consequences, and that none of them makes the question about the occasion unnecessary.

Fazzano and Gallo (2015) criticize explanations that treat violence as a defense against the aggressor's own impulses, or as a defense of the ego, because they attribute the cause to an "inaccessible and immutable" entity, disregard the contexts the individual has gone through and, in their words, free social agencies from acting. Mizael (2018) takes up the point when she proposes reducing the essentialisms that treat homophobia as an entity inside the individual. Functional analysis changes the question: from what lies inside the one who speaks to the conditions under which the speaking occurs (Skinner, 1953/2003).

CONCEPT IN USE

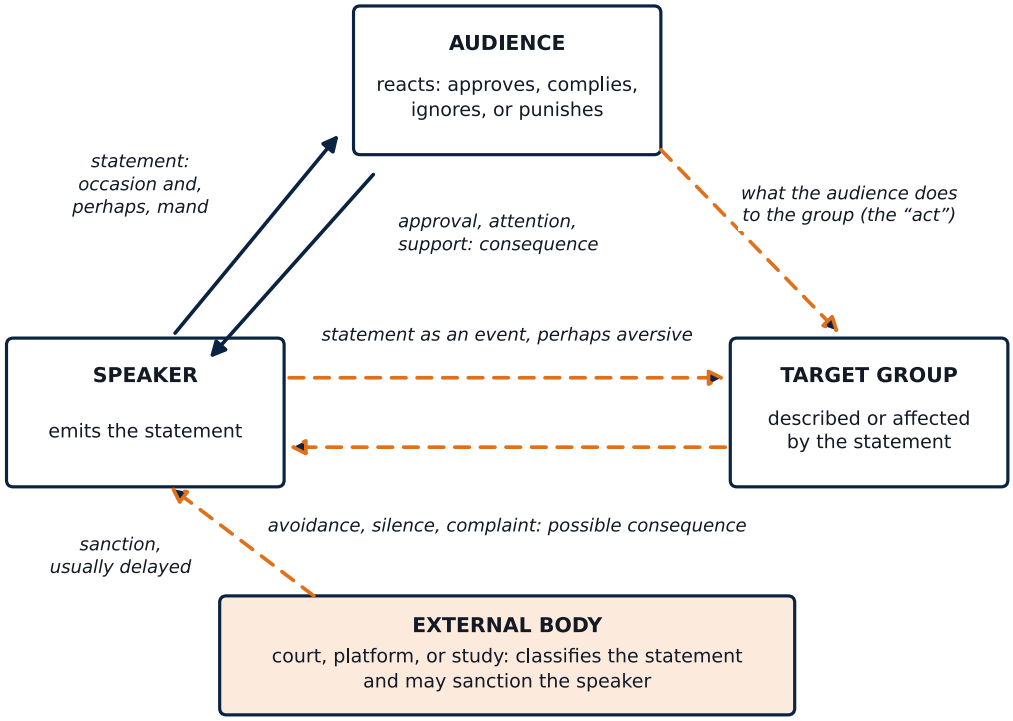
Interlocking behavioral contingencies (Glenn, 2004)

Operant contingencies are interlocked when the behavior of each of two or more people serves as an environmental event for the behavior of the others (Glenn, 2004). The outcome of interlocked behavior is not the cumulative effect of each person behaving alone, but rather the effect of their interrelated behavior, "more than" or "different than" what each would produce in isolation. In this chapter, the interlocking has three participants: the speaker, whose response is the statement; the audience, whose reaction is a consequence for the speaker's behavior; and the target group, for whom the statement is an event and whose reaction may also be a consequence for the speaker's behavior. The application to these three is made here, not by Glenn.

Figure 5.1 brings the relations together. The two solid lines, between speaker and audience, are those for which there are some indirect data, discussed in Section 5.5: leaders' speech that precedes changes in what their supporters do, and public expression that tracks the perceived popularity of what is said. The dashed lines are those the chapter proposes and no verified source measures: what the statement does to the target group, what the target group does in return, and what the audience does to the group. The external body is not part of the contingency at the moment of emission: it classifies afterward and, when it sanctions, does so with a delay. It will belong to the contingency only in the case of recurrence and of a suppressive effect on subsequent responses of the same class. In the opening case, the documented consequence for the speaker came from a court, years later, and not from the show's audience.

FIGURE 5.1

Speaker, audience, and target group: interlocking contingencies around a statement



Solid line: relation with indirect data in a verified source (Section 5.5). Dashed: conceptual relation or hypothesis.

Note. Original diagram by the author, based on Glenn (2004), Skinner (1957), and Fazzano and Gallo (2015). It does not represent data.

5.2 The disguised mand: the opinion that functions as an instruction

The mand, in the vocabulary of Chapter 3, is the verbal operant whose response is reinforced by a characteristic consequence and which is therefore under the control of deprivation or aversive stimulation (Skinner, 1957). Skinner distinguishes kinds of mands by the consequences provided to the listener's behavior when the listener complies with the speaker's mand. In the request, the listener is already disposed to give what is asked, and the response merely indicates what the speaker will accept. In the command, the speaker's response sets up an aversive situation that the listener can escape only by supplying the appropriate mediation: "Hands up!" does more than specify an action; it is a threat the victim escapes by raising his hands, and "Your money or your life!" contains both parts, the one that specifies the reinforcement and the one that names the aversive consequence. Skinner writes that military orders are obeyed by virtue of a kind of permanent threat. When the listener

gains something of his own, the mand is advice, if there is positive reinforcement, or a warning, if there is escape from aversive stimulation; permission cancels a threat that was holding the listener back.

CONCEPT IN USE

Disguised mand (Skinner, 1957; Guerin, 1994)

Skinner uses the expression in two passages. In the first, concerning negative audiences: verbal behavior about the listener, or about people important to him, may be received as aversive, and with certain listeners we learn to steer clear of mands or to use disguised mands in their place. In the second, the call "Call for Philip Morris" functions as a mand although it is disguised as a bellhop paging someone. Skinner does not develop the concept beyond that. Guerin (1994) brings it closer to what this chapter needs: attitude statements can result in specific consequences and can thus qualify as mands; a boss's remark that she dislikes people drinking coffee while working is not an innocent tact of a private attitude, and it works as if she had manded "Do not drink coffee while you work or else!", whether or not it was intended. Medeiros (2013) defines disguised mands as verbal responses with the topography of a tact or intraverbal whose occurrence is controlled by the reinforcers that maintain mands, even though their topography does not clearly make the reinforcer explicit. In this book, a *disguised mand* is a statement with the form of a tact or an opinion whose function, on the available evidence, is that of a mand. The application to political speech is a hypothesis proposed here.

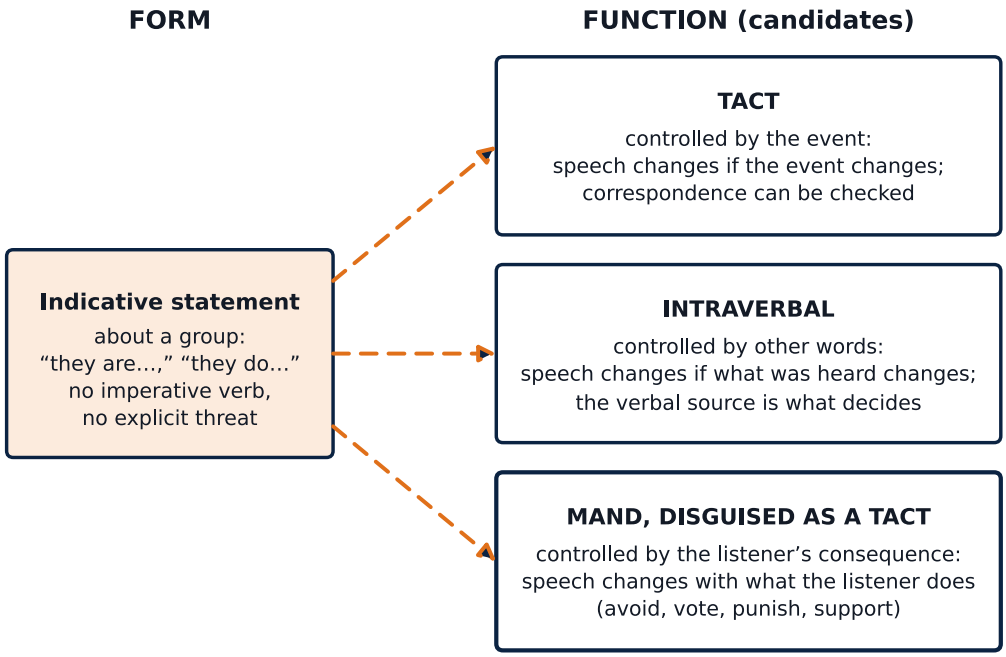
The two criteria Skinner's text provides are, therefore, a condition of appearance and a criterion of identification. The condition is an audience that punishes the explicit mand: with such an audience, the speaker comes to use the form that is not punished. The criterion is the divergence between form and function: the form is that of something else, the function is that of a mand. This chapter takes only these two and Guerin's reading; the rest is application.

The application is this. A statement such as "they are dangerous" has the form of a tact, in the indicative mood, about a group. If its function is that of a mand to the audience, it will be controlled by what the listener does afterward, and what is being asked is: avoid them, vote against them, support those who attack them, or at least do not punish. Nothing in the form says this. Figure 5.2 shows the three functions among which the same form may lie, and which Section 3.3 has already distinguished: tact, intraverbal, and mand, that is, the three verbal operants whose topography is not determined by a point-to-point relation with a verbal antecedent stimulus.

What separates the mand hypothesis from its rivals are the three questions the Chapter 3 analysis form already asked, now applied to the listener's behavior. What the listener does afterward: an observable response, not an intention attributed to the listener. What the speaker does when the listener does not comply: Skinner notes that repetition may signal nothing more than excessive response strength and that, when verbal behavior fails to change the conditions that give it strength, one response may give way to another, so that, if straight repetition is punished, the form varies (Section 3.5). And whether the form changes when the audience changes: with an audience that punishes the explicit mand, disguise is to be expected; with an audience that reinforces it, the explicit mand is what one would expect. As for the assertive autoclitic, which "enjoins the listener to accept a given state of affairs" and which Section 3.6 discussed, Skinner already classifies it as a special sort of mand (p. 326); "never," "not ever," and "always" therefore carry an embedded mand that does not depend on the subject matter.

FIGURE 5.2

The disguised mand: the same opinion form, three candidate functions



The form is the same; what sets it apart is what controls it (Skinner, 1957, p. 36).

Note. Original diagram by the author, based on Skinner (1957) and Guerin (1994). It does not represent data. The three arrows are dashed because the three functions are hypotheses for any particular statement.

None of the verified sources records all three things for a statement classified as hateful. That is why the interpretation as a disguised mand is a possibility to be investigated and not a classification of the statement, and that is why, in the chapter's boxes, it stays in the hypothesis column.

5.3 Aversive control directed at a group

Skinner (1957) presents the aversive control of verbal behavior through the reduction of aversive stimulation. When the aversive stimulus itself is reduced, the behavior is escape; when some condition that characteristically precedes it is reduced, it is avoidance. His verbal examples are "Stop it!", reinforced when it puts an end to physical injury, and "Don't touch me!", reinforced when it puts an end to the threat of that injury. The point that matters here is that aversive stimulation is not only physical. Skinner writes that one man can stimulate another aversively by many means, among them "many sorts of 'verbal damage'" (p. 54), and that the threat that controls verbal behavior does not necessarily imply violence: there are mild forms, and a question carries a slight generalized threat, in the sense that not answering is followed by censure.

CONCEPT IN USE

Aversive control (Skinner, 1957; Fazzano and Gallo, 2015)

Aversive control is the control of behavior by aversive stimulation: behavior that removes or avoids it is reinforced (escape and avoidance), and behavior that produces it tends to be suppressed (punishment). Fazzano and Gallo (2015) define punishers, following Sidman (1989/2003), as the production of something that would normally be a negative reinforcer or the withdrawal of something that would normally be a positive reinforcer. In the case they analyze, the aversiveness is not in the stimulus: it is learned. Their analysis isolates a three-term contingency, which they acknowledge to be a simplification of the contingencies involved in aggression (note 3). Aversive stimulation can be verbal (Skinner, 1957) and can be mild.

Fazzano and Gallo's (2015) article is the Brazilian behavior-analytic account of homophobia among the verified sources, and it is, by its own declaration, a functional hypothesis and not a functional analysis: the hypothesis is formulated, and no experiment is run, for ethical reasons. In their terms, homophobia is a set of complex behaviors, operants and emotional responses, of physical, psychological, or sexual aggression against homosexuals or those who identify with homosexual culture, and it is related to aversive control. The part that interests this chapter is the one that deals with maintenance. For the aggressor, the immediate consequences are the removal of aversive stimula-

tion, which the authors assume to be learned, and, perhaps, the group's approval. Punitive consequences, such as criminal proceedings, fines, or imprisonment, are distant in time, and the aggressor may never come into contact with them, because, according to the authors, in 2015 there were no aggravating circumstances for homophobia in the criminal code and the number of reported complaints was small. The text is from 2015, and the legal situation may have changed; no claim is made here about the current state of the law. At the same time, for us to regard them as punitive consequences, the recurrence of the behavior–consequence relation would have to be observed. That is not the case.

The same analysis has a victim's side, and it is the side least developed by studies of speech. Fazzano and Gallo (2015) treat aggression, from the standpoint of the one who receives it, as punishment of the behavior that was being emitted: two men attacked for walking arm in arm will probably not embrace again in the same place or in similar contexts, and the punishment generalizes, so that the person may stop going to the place of the aggression and, in some cases, stop leaving home, becoming isolated. Staying at home, avoiding relationships, or staying "in the closet" would be incompatible responses that prevent further aggressive stimulation. This is an analysis of the effects of aggression, not of statements, and the chapter includes it because the hateful statement, if it is aversive to the target group, has this side: the consequence such a statement produces in the group is a reduction in what the group does in public. This analysis corresponds to the second case pointed out by Skinner, that of aversiveness. If a disguised mand results in aggression and in the suppression of the behavior of those attacked, the suppression negatively reinforces the behavior of emitting the disguised mand. No verified source measures this for statements.

There is a methodological caveat from the same authors that applies to the whole book. Behavior is defined as violence or aggression by its function in given contexts: two friends who address each other with a generally pejorative term may simply be expressing rapport and intimacy, and the same term would constitute verbal aggression if it produced sadness, embarrassment, or humiliation in the victim. This agrees with this book's labeling rule. The word does not decide: what decides is what it does, to whom, and on what occasion.

Two gaps complete the picture. The first is the one Mizael (2018) documents in a literature review. The author searched for the terms "homosexuality," "sexual orientation," "heterosexuality," and "bisexuality" (and their equivalents in English and Spanish) in 13 behavior analysis journals, three Brazilian and 10 international, with no time limit, and included only the articles in which homosexuality was a central topic. She found three, and a supplementary search

in SciELO, in Portuguese, which added the term "behavior analysis," yielded the fourth. One of them is Fazzano and Gallo (2015), the only one published in a psychology journal not specific to the field. In general, the four analyzed homosexuality or homophobia in terms of the contingencies of reinforcement and punishment to which people may have been exposed, or examined possible biological causes of homosexuality. What was missing, according to the author, were functional analyses of behaviors considered prejudiced and the study of jokes and other potentially discriminatory behaviors as practices reinforced by the verbal community. The second gap is the one Fazzano and Gallo (2015) acknowledge. The Brazilian data they cite come from Disque 100 (Dial 100), a complaints hotline on which the Secretariat of Human Rights based itself, in 2011, to characterize homophobic violence: 6,809 reports of violations against lesbians, gays, bisexuals, and transsexuals, involving 1,713 victims and 2,275 suspects. In 61.9% of the cases the victims knew the aggressor, and the most frequent forms of violence were psychological violence (42.5%), discrimination (22.3%), and physical violence (15.9%). Since these are only complaints, the authors themselves write that the numbers may not meaningfully express the Brazilian reality.

5.4 The group label and the transfer of aversive function

Chapter 4 showed how a word acquires an evaluative function without direct contact with its referent: by belonging to an equivalence class or by entering into relations of opposition and comparison with others (Sections 4.2, 4.3, and 4.6). For speech about groups, this means that the statement does not require the listener's prior direct contact with the group for an aversive function to transfer to the group; it is enough that the group's label be in a class or a relation with stimuli that have an aversive function. Fazzano and Gallo (2015) make the point in other terms: if the presence of homosexuals is not aversive in itself, the aversiveness was learned, and the learning may occur through shaping, instruction following, and imitation, and instruction following is a verbal process. If a sentence with an assertive autoclitic describes a group in pejorative terms ("Women vote badly!"), then, even if the listener has never come into direct contact with aversive stimuli resulting from women's votes, the possibility of women voting acquires an aversive function by transfer of function or transitivity.

Mizael (2018) gives an example of what transfer means for sexual prejudice, and gives it as a theoretical illustration and research proposal, not as data: a boy may learn that "viado" (a Brazilian Portuguese slur for gay men) is something bad without knowing what the word means; when he grows up, he learns that the word is equivalent to "gay" or "homosexual," and the negative

attributes pass to those who define themselves, or are perceived, as gay. The mechanism the example describes is the one Chapter 4 showed in the laboratory with other stimuli. Dixon et al. (2003), in the same vein, analyze prejudice and acts of terror as a product of the transformation of functions of individuals or groups on the basis of frequently indirect contact with a few of their members, as an argument and without data (Chapter 4, Section 4.6).

What this adds to the chapter is a third position for the speaker, besides object and agent of aversion: that of the one who establishes the relations. A statement that associates a group's label with terms that have an aversive function, or that opposes it to an "us" with a positive function, would be doing for the listener the work that the laboratory procedure does for the participant. As data, here, there is only the procedure; the hypothesis that political speech does this, and with what effect on what the listener says and does afterward, is proposed here, and testing it would require dated statements and measures of the listener before and after, not just analysis of the text.

5.5 From speech to act: what the data allow us to state

The passage from speech to act combines two questions that are best kept apart. The first is whether the statement changes what the audience does, for example, acting against the group or no longer punishing those who act. The second is whether the statement changes what the speaker or other speakers say next. The verified studies are, for the most part, about the second; those about the first do not deal with hate speech. Table 5.1 brings them together.

TABLE 5.1

What six verified studies measure as "speech," as "act," and what they allow us to state

Study	Measure of "speech"	Measure of "act" or effect	What it allows	What it does not allow
Ajzenman et al. (2023, working paper version)	The president's participation in a demonstration (3/15/2020) and an official address (3/24/2020) against social distancing; not speech about a group	Cell-phone-based distancing index, by municipality	Larger drop in distancing in municipalities where Bolsonaro had more than 50% of the vote in the first round of 2018, about 1 percentage point on average, with no effects before the events	Separating conduct from speech; speaking of hate speech; the mechanism (media, evangelicals) is only suggestive

Study	Measure of "speech"	Measure of "act" or effect	What it allows	What it does not allow
Bursztyn et al. (2020)	Random information about who won in the area (perceived popularity of Trump)	Donation to an anti-immigration organization, in public or in private	Public donation, but not private donation, varied with perceived popularity	Effect of speech; it only manipulates the perception of popularity; United States
Álvarez-Benjumea and Winter (2018)	Comments in an experimental forum, scored for hostility	The comment itself	Removing prior hate content reduced hostile comments (5.88% to 0.99%); counterspeech had no significant effect	Generalizing; paid, anonymous participants
Siegel et al. (2021)	Tweets classified by dictionary and by Reddit communities	Aggregate level of speech, with no measure of acts	No persistent increase in hate speech during the campaign or after the 2016 election	Ruling out other platforms; conservative dictionary
Marino-Fages and Martínez-Pozo (2026, working paper)	Geolocated tweets classified by a model and by dictionary	Level of speech, by municipality	Increase after the 2018 election, larger where Bolsonaro had fewer votes, in homophobia and sexism	Measuring perceived social cost; the mechanism is the authors' reading
Hangartner et al. (2021)	Xenophobic tweets from 1,350 users	Retrospective deletion and prospective creation of hate speech	Empathy-based counterspeech increased deletion and reduced creation; humor and warning of consequences, no consistent effect	Effect on those who only read

Note. Figures from the articles. Ajzenman et al. was read in the 2020 working paper version, from Universidad Torcuato Di Tella, without the journal's pagination. Marino-Fages and Martínez-Pozo is a working paper that, according to its cover, has not undergone peer review. The studies by Bursztyn et al., Álvarez-Benjumea and Winter, Siegel et al., and Marino-Fages and Martínez-Pozo are described in Chapter 10 (Sections 10.3, 10.4, and 10.6); that of Hangartner et al. is discussed in Chapter 12.

The study by Ajzenman et al. (2023) is the only one, among those verified, that links the speech of a Brazilian head of state to what his audience does, and what it measures is distancing during the pandemic, not acts against a group. The authors hypothesized that the president's supporters would be far more easily persuaded by his speeches than his critics would be, even in high-cost decisions, such as whether or not to follow the preventive measures recommended by the World Health Organization and by subnational governments. The outcome was a distancing index calculated each day, for each municipal-

ity, from the location of about 60 million anonymized cell phones: the share of devices that, on that day, remained no farther than 450 meters from what was identified as their home. The index ranges from 0 to 1, had a mean of 0.37 over the period (from February to the first two weeks of April 2020), and is available for 3,975 of the 5,571 municipalities, because small ones do not have enough devices (Section II). The two events studied were the president's participation in the demonstrations of March 15, 2020, and the official address of March 24, in which he pushed for an end to the distancing measures of subnational governments.

The comparison is between pro-government municipalities, where Bolsonaro had more than 50% of the vote in the first round of 2018, and the others, day by day, before and after the two events. The authors call the design an event study, with day and municipality fixed effects (which net out what is specific to each municipality and what changes in all of them on the same day) and separate estimates for each day before (leads) and after (lags) the event (Sections I and V.1). In the days before, with one exception, the difference between the two groups is indistinguishable from zero, that is, there was no prior difference between them. From the second day after the events, distancing was significantly lower in pro-government municipalities, about 1 percentage point on average. In a specification restricted to the states in which he had more than 50% of the vote in at least a third of the municipalities, comparing, within each state, municipalities above and below the median vote, the effect comes close to 2 percentage points, against a mean of 0.37 for the period (V.2). Using the coefficients and the population of pro-government municipalities (about 100 million), the authors calculate, as a rough estimate, about 1 million more people straying further than 450 meters from home on each of the ten following days. They call the result a "strong persuasion effect," and say they cannot present conclusive results about the mechanisms (V.3 and VI).

In operant terms, the study is compatible with a reading of the address as a rule, or as a mand, for those who already had a history of reinforcement with the president, and it is also compatible with imitation of his conduct at the demonstration. The authors estimate the joint effect of his participation in the demonstrations and of his messages against distancing (Section V.1) and, therefore, do not distinguish between the two. What this chapter gains is methodological: measuring what the audience does after a speech is feasible on a large scale for behavior that cell phones record. The question of acts against a group, by contrast, would require a different measure. The condition for reading it is that the effect appears where the audience already supported the leader. Baldwin and Baldwin (1986) hold that aspects of the observer's rela-

tion to the model (admiration, respect, and affection) are fundamental elements in the probability that the observer will behave like the model. These are not data on hate speech, and should not be read as such.

The next four studies, described at greater length in Chapter 10, deal with what people say or agree to do in public after a piece of information or an event, and two of them are experiments. Bursztyn et al. (2020) recruited 1,600 residents of the Pittsburgh area from an online panel and informed, at random, half of them that Trump had won in the metropolitan area and the other half that Clinton had won in the county where the city is located. They then offered a bonus to those who authorized a donation to a strongly anti-immigration organization, and the decision was either anonymous (private condition) or visible to peers in the area (public condition). In public, 33% of those told of Trump's victory and 20% of those told of Clinton's agreed to donate; in private, 31% and 30%. The difference between these differences was significant: the effect of the information appeared in public and not in private.

Álvarez-Benjumea and Winter (2018) recruited 180 participants on a crowdsourcing platform and asked them to comment on news items in an experimental forum, under four conditions: the baseline, with friendly, neutral, and hostile comments; the censored condition, with prior hate content deleted; the extremely censored condition, with friendly comments only; and the counterspeech condition, with the hostile comments accompanied by replies emphasizing their unacceptability. Hostile comments were 5.88% at baseline and 0.99% in the censored condition, and only that reduction was significant; counterspeech (4.44%) had no significant effect.

The other two are observational studies of texts on social media. Siegel et al. (2021) analyzed more than 750 million tweets about the 2016 U.S. election and almost 400 million from a random sample of American users and, with the two measures they adopted (a dictionary augmented by machine learning and a classification based on Reddit communities linked to the alt-right), found no persistent increase in hate speech during the campaign or after the election; there were only brief spikes.

Marino-Fages and Martínez-Pozo (2026) gathered more than 37.6 million tweets in Portuguese, geolocated in Brazil, from July 2017 to December 2019, and compared the period before and after the 2018 election according to Bolsonaro's share of the first-round vote in each municipality: hate speech increased after the election, more where he had fewer votes, and the increase was concentrated in homophobia and sexism. The four measures and the four designs differ, and the studies neither confirm nor refute one another. Hangartner et al. (2021), in Chapter 12, complete the list on the intervention

side: they randomly assigned Twitter users who had authored xenophobic tweets to receive or not receive counterspeech, and empathy-based counterspeech increased deletion of the original tweet and reduced the creation of new xenophobic tweets, with caveats that Chapter 12 details.

Álvarez-Benjumea and Winter (2018) draw from the experiment a conclusion relevant to the operant reading: normative behavior online would be motivated more by descriptive norms (what others do) than by injunctive norms (what one ought to do). In the vocabulary used here, the descriptive norm is what the audience does in the presence of the statement, a discriminative stimulus for whoever speaks next; the injunctive norm is what the community says one ought to do, a rule (Section 3.7). The counterspeech condition presents the injunctive norm, with replies stating that hate is unacceptable, and had no significant effect; the censored condition changes what one sees others say and had an effect. The authors suggest, to account for the failure of counterspeech, that replies that are themselves hostile add ambiguity about the norm. The experiment does not separate the two readings and is a single experiment with paid, anonymous participants; the reading is compatible with the view that, for the speaker, what the audience does weighs more than what it says one ought to do.

Missing from Table 5.1 is the column the section's title called for: that of acts against the group. Among the verified sources, none links hateful statements to acts of violence against the target group, in Brazil or elsewhere. Studies that do so exist in the literature and are not cited here because their full text could not be read through legal means. The gap is, therefore, one of the body of sources read and should be read as such: it is neither an absence of a relation nor a demonstration of one.

DATA	In municipalities where the president had more votes, social distancing fell after two events in which he publicly opposed distancing, by about 1 percentage point on average, with no effect before the events (Ajzenman et al., 2023). Public expression of an anti-immigration position, but not private expression, tracked the candidate's perceived popularity (Bursztyn et al., 2020). In Brazil, hate speech in geolocated tweets increased after the 2018 election, more where Bolsonaro had fewer votes (Marino-Fages and Martínez-Pozo, 2026).

INTERPRETATION	What the audience does and what others say seem to depend on what the audience believes others accept and on what the leader does and says. The evidence is indirect, from different designs and measures, and none of it measures a politician's speech in a way that would allow the effect to be attributed to the content rather than to the conduct, the office, or the moment.
HYPOTHESIS	That a politician's hateful statement functions as a mand to the audience, with a measurable effect on what the audience does to the target group, is a hypothesis, and no source tests it. Testing it would require dated statements, a measure of what the audience does afterward, and comparison with those who were not exposed and, for the target group, measures of what it stops doing in public.

5.6 A case with external classification: the 2011 remark and what came after

The opening case is the only one among the verified sources with the three elements that the chapter's rule requires in order to speak of hate speech without labeling on our own: the transcribed statement, the date and the outlet, and classification by an external body, named and dated. What it lacks is what the Chapter 3 analysis form asks for under audience present and candidate controlling variables, the composition of the audience and its reaction, which remain blank; the section says which ones.

The documented fields are these. The remark: "That doesn't even cross my mind, because they had a good upbringing. I am a present father, so I don't run that risk," in response to a question about what he would do if he had a gay son (Agência Lupa, 2026). The date and the outlet: 2011, the TV show CQC, on the Bandeirantes network; the month is not given in the source. The external classification: a conviction for collective moral damages, upheld by the TJRJ (the Rio de Janeiro State Court of Justice) in November 2017, following statements with homophobic content made on the show, with a penalty of R$150,000 set in 2015, all according to Lupa; the court's text was not read. Also according to Lupa, the case was still pending in 2018, and there are other statements by the deputy with content it calls homophobic, in a debate on TV Câmara (the Chamber of Deputies' channel) and in an interview with *Playboy*, in 2011. On August 3, 2018, in the same fact-check, the sentence "I have never been homophobic," said on Central das Eleições, a GloboNews program, was rated by Lupa as "False" (Chapter 3, Section 3.8, case 4).

The fields of the analysis form that remain blank are the most important for operant analysis: the reaction of the show's audience, its composition, what the speaker did afterward with different audiences, and any measure of what the target group did or stopped doing. None of this is in the sources, and the chapter does not fill them in by inference. The 2018 denial was classified by Lupa and appears among the cases in Chapter 3, where it received the same treatment as Lula's remarks; the analysis there is of the form of the sentence and of its correspondence with the data, and the analysis here is of the relation between the 2011 remark, the sanction, and the 2018 remark.

A caveat about symmetry, in the same spirit as the one Chapter 3 made. The verified sources do not include, for another political camp, an equivalent case, that is, one with a statement transcribed, dated, and classified by a named external body. This is a limit of the body of sources read, not a property of the political camps. The analysis form and the criterion would be the same, and the sampling rule of Section 2.9 is what would make it possible to find such cases.

DATA	An answer, given in 2011 on a TV show to a question about what the deputy would do if he had a gay son, led to a conviction for collective moral damages upheld in 2017, according to Lupa. In 2018, the same politician, in an interview, said he had never been homophobic, and Lupa rated the sentence as "False."
INTERPRETATION	The answer was emitted in response to a question, and the statement rests, in its form, on two descriptions of the speaker himself ("I am a present father," "they had a good upbringing") and on a word with an evaluative function ("risk"), which makes having a gay son an occasion to be avoided. From the form alone, one cannot decide whether the statement is a tact about what the speaker would do, a learned intraverbal, or a mand to the audience (Sections 3.3 and 5.2). The conviction is an aversive consequence, delayed by years, and it comes from a legal body, not from the show's audience. The conviction for collective moral damages is a legal classification of the effect on the group; it is not a measure of the target group's behavior.
HYPOTHESIS	That the 2011 remark was reinforced by the show's audience (attention, approval, support) is a hypothesis, and the source records nothing about the audience's reaction. That the conviction altered the emission of statements of this kind, and that the 2018 denial ("I have never been homophobic") is a function of this history, whether as avoidance of punishment or as a tact distorted by the listener's consequence (Section 3.4), is a hypothesis and would require a series of dated statements, before and after 2015 and 2017, with their audiences.

5.7 Rival interpretations and limits

The chapter's reading, that the hateful statement is maintained by social and aversive consequences and may function as a disguised mand, has rivals, and Table 5.2 brings them together with what would separate them from it.

The chapter's limits are those of its sources. On aggression and hate speech, the body of sources read includes one article offering a functional hypothesis about homophobia (Fazzano & Gallo, 2015) and one review showing how scarce the field is (Mizael, 2018); the experiments of Chapter 4 deal with prejudice, not with hateful statements, and there is no study testing the mand function or aversive control for statements. The evidence on speech and acts comes from diverse designs and measures, from a single country, a single platform, or an experimental forum, and none of it concerns a politician's speech with an identified audience. The only case with external classification

comes from a single political camp, and the classification is that of one body and of Lupa's accounts, recorded with attribution. The dashed relations in Figure 5.1 have no measure among the sources. Section 5.5 says which gaps belong to the body of sources read, and not to the literature.

TABLE 5.2

Rival interpretations for a statement classified as hateful and what would separate them from the main operant reading

Rival	What it says	What would separate it
Emotional state	The statement is an effect of a state of anger or aversion, more than of social consequence; Skinner notes that making a man angry increases the probability of abusive or aggressive behavior (1957)	Whether emission varies with the presence, absence, and identity of the audience; whether it stops when no one is there
Group identity and ritual	The exchange is the reinforcer, and the content varies without altering the effect (Guerin, 1994)	Whether the content changes without changing the reaction; whether the reaction is to adherence and not to what is said
Intraverbal	It repeats what was heard in the family, the religion, or the group, through modeling and rules (Fazzano and Gallo, 2015)	Acquisition history; whether it varies with the verbal source; whether it is sensitive or insensitive to consequences
Joking	The function is rapport or attention, not aggression (Fazzano & Gallo, 2015; Mizael, 2018)	The effect on the target group and on the audience; whether the term produces embarrassment or humiliation in those described
Tact under the control of an event	The sentence is a report of an event, with checkable correspondence	Checking of correspondence by an agency; whether it changes with the event
Artifact of classification	Another body, with another criterion, would not classify it this way (the classifiers and dictionaries used in the studies make errors: accuracy of 77 to 79% in Marino-Fages and Martínez-Pozo, 2026)	Comparing criteria and bodies; reclassifying a sample by more than one criterion

Summary

A statement classified as hateful is a label from an external body, not an explanation. Operant analysis asks what consequences maintain it, what function it has for the audience, and what effect it has on the group it refers to. For the speaker, the candidates are the audience's approval, the removal of learned aversive stimulation, and functions that vary with the occasion. For

the audience, the hypothesis is that a statement with the form of a tact operates as a disguised mand: condition of appearance, an audience that punishes the explicit mand; criterion, the form of one thing, the function of another. For the target group, the stimulation is aversive and, in Fazzano and Gallo's (2015) analysis of aggression, punishes what the group used to do in public and reinforces what hides it.

The verified data do not link hateful statements to acts against the group. They show, with measures and designs that do not add up, that leaders' speech precedes changes in what their supporters do (in social distancing, not in hate), that public expression tracks perceived popularity, that moderate removal of hostile content reduces hostility in an experimental forum, and that, in Brazil, hate speech in tweets increased after the 2018 election, more where the winner had fewer votes. The only case with external classification in the sources is the 2011 remark and the conviction upheld in 2017, according to Lupa; the fields of the analysis form that matter for operant analysis are blank, and they stay that way.

Chapter 6 takes up aversive control from the side of those who receive it, as countercontrol (Section 6.4); Chapter 10 defines an audience's window by what it punishes and what it does not punish and describes the studies in Table 5.1 at greater length; Chapter 12 deals with counterspeech and the other interventions on the speaker and on the audience.

References

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Ajzenman, N., Cavalcanti, T., & Da Mata, D. (2023). More than words: Leaders' speech and risky behavior during a pandemic. *American Economic Journal: Economic Policy*, *15*(3), 351–371. https://doi.org/10.1257/pol.20210284

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Baldwin, J. D., & Baldwin, J. I. (1986). *Behavior principles in everyday life*. Prentice-Hall.

Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, *110*(11), 3522–3548. https://doi.org/10.1257/aer.20171175

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues*, *12*(2), 129–147. https://doi.org/10.5210/bsi.v12i2.40

Fazzano, L. H., & Gallo, A. E. (2015). Uma análise da homofobia sob a perspectiva da Análise do Comportamento [An analysis of homophobia from the perspective of behavior analysis]. *Temas em Psicologia*, *23*(3), 535–545. https://doi.org/10.9788/tp2015.3-02

Glenn, S. S. (2004). Individual behavior, culture, and social change. *The Behavior Analyst*, *27*(2), 133–151. https://doi.org/10.1007/bf03393175

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Marino-Fages, D., & Martínez-Pozo, A. (2026). *Hate in the Tropics: Political leaders and the social acceptability of online hate speech* (Documento de trabajo N.º 391). RedNIE. https://rednie.eco.unc.edu.ar/files/DT/391.pdf

Medeiros, C. A. (2013). Mentiras, indiretas, desculpas e racionalizações: Manipulações e imprecisões do comportamento verbal [Lies, innuendos, excuses, and rationalizations: Manipulations and imprecisions of verbal behavior]. In Associação Brasileira de Psicologia e Medicina Comportamental, *Comportamento em foco* (Vol. 2, pp. 157–170). ABPMC.

Mizael, T. M. (2018). Perspectivas analítico-comportamentais sobre a homossexualidade: análise da produção científica [Behavior-analytic perspectives on homosexuality: An analysis of the scientific literature]. *Revista Perspectivas*, *9*(1), 15–28. https://doi.org/10.18761/PAC.2017.011

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 U.S. election campaign and its aftermath. *Quarterly Journal of Political Science*, *16*(1), 71–104. https://doi.org/10.1561/100.00019045

Skinner, B. F. (2003). *Ciência e comportamento humano* [Science and human behavior] (J. C. Todorov & R. Azzi, Trans.; 10th ed.). Martins Fontes. (Original work published 1953)

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

instituto WALDEN4 | ACADEMIA WALDEN4



EVIDENCE-BASED PRACTICE

How a Practice Becomes Evidence, and What the Evidence Does Not Say

The complete module on evidence-based practice: designs, criteria, levels of evidence, and the limits of each. The same question this book asks of the data, asked of clinical practice.

Free access, no sign-up required

walden4.com.br/pratica-baseada-em-evidencias

CHAPTER

6

Digital Bubbles as Verbal Communities and the Avoidance of Countercontrol

Nyhan et al. (2023) tested, in a field experiment on Facebook during the 2020 U.S. presidential election, what happens to attitudes when exposure to like-minded political sources is reduced. The participants were 23,377 users in the United States. For those assigned to treatment, the authors demoted in the feed, from September 24 to December 23, 2020, content from friends, groups, and pages expected to have the same political leaning as the participant; the control group's feed was not systematically altered. The share of what users saw coming from these sources fell from 53.7% (control) to 36.2% (treatment), about one third less. The authors then measured eight attitudes they had registered in advance (preregistration), such as affective polarization, ideological extremism, evaluation of candidates, and belief in false claims, and found no measurable effect on any of them: they rule out effects of 0.12 standard deviations or more, that is, of little more than one tenth of the typical variation of each measure. The exposure data in the same article, from 2020, which cover the active adult Facebook population in the United States, show that content from like-minded sources makes up the majority of what users see. If homogeneous exposure were what sustains attitudes, cutting a third of it should have moved something, and it did not.

The result admits two readings. In the first, the bubble does not have the effect attributed to it. In the second, what was reduced was not what matters: the literature measures the bubble by exposure, that is, by what reaches the feed, whereas the analysis of verbal behavior asks about something else, about who reinforces what the speaker says and about how much of what comes from outside the community that reinforces the speaker actually

reaches him or her. This chapter takes the second reading as a working hypothesis and states, at each step, what supports it. The first will be presented, with its arguments, in Section 6.8.

The chapter has three parts. Sections 6.1 to 6.3 describe the bubble as a verbal community and confront it with the data on echo chambers. Section 6.4 deals with what protects the speaker from correction, which is the central proposal of the chapter and the part with the least empirical grounding. Sections 6.5 to 6.7 take it to the Brazilian material, including an exploratory analysis of two political videos on Instagram (Section 6.6), and to a method by the author. Throughout, data, interpretation, and hypothesis are kept separate, and calling a repertoire avoidance of countercontrol is always interpretation.

6.1 The bubble as a verbal community: audience and source of reinforcement

Guerin (1992) notes that Skinner's verbal community, on a smaller scale, is equivalent to Durkheim's collectives and to Mead's generalized others, and that changes in language use may be under the control of smaller verbal subcommunities. It is in this sense that this chapter speaks of a bubble: not a place or a distribution of content, but a set of listeners with a shared history of reinforcement, who reinforce some things that are said and not others.

CONCEPT IN USE

Verbal subcommunity (Guerin, 1992), revisited from Chapter 3

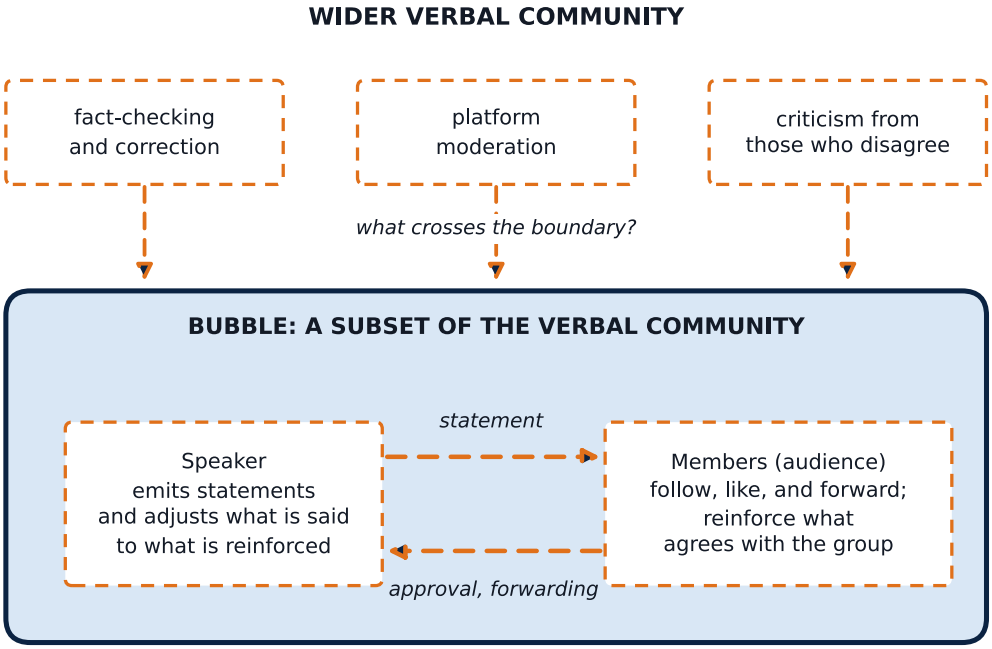
Chapter 3 (Section 3.4) presented the distorted tact: the tact whose control by the event has been displaced by special consequences provided by the listener (Skinner, 1957). What this chapter adds is scale. For Guerin (1992), control over what is said may lie in the hands of verbal subcommunities smaller than the wider verbal community: if a small group reinforces certain reports and punishes others, the characteristics of the group come to control what is reported, even though the consequences are generalized, and Skinner (1957, as cited in Guerin, 1992) called these tacts distorted. Section 6.3 presents the three routes by which, according to Guerin, what is said becomes detached from the physical environment and comes to be controlled by the social environment.

Chapter 3 presented the two terms the description needs: the audience is an occasion for what is said (Section 3.1), and its reaction is a consequence (Section 3.3; Guerin, 1994). What remains is to say what changes when the audience is a subcommunity. What changes is the composition of the reactions. If the listeners who react share a history, the reactions the speaker receives tend to come from the same direction, and what the wider community would say

about the statement (correct it, criticize it, refuse to go along with it) has little effect or arrives by other routes. The chapter's hypothesis is that this, and not mere exposure, is what makes the bubble a distinct environment of verbal control.

FIGURE 6.1

The bubble as a verbal subcommunity: what the studies measure and what the model assumes



Solid line: relation supported by data (homophily of interactions, Cinelli et al., 2021; like-minded sources in the feed, Nyhan et al., 2023). Dashed: hypothesis about who reinforces what and how much external correction reaches the speaker; no study in the chapter measures this.

Note. Original diagram by the author, without data. Solid line: relation supported by data (homophily of interactions, Cinelli et al., 2021; predominance of like-minded sources in the feed, Nyhan et al., 2023). Dashed: hypothesis. The relations inside the bubble (who reinforces what) and how much of the external correction reaches the speaker were not measured in any study in the chapter.

The literature shows homophily in interactions and a predominance of like-minded sources in the feed (Section 6.2). The internal arrows, in which the speaker adjusts what is said to what is reinforced and the audience reinforces what agrees with the group, are supported by Guerin's analysis and, for opinions, by the 1956–1969 experiments discussed in Chapter 9, but they have not been measured in digital bubbles. And the question the figure asks at the boundary, how much of the correction directed at what the speaker said reaches him or her, has not been measured either: Nyhan et al. (2023) measure the share of content from cross-cutting sources in the feed, which is not the same thing.

6.2 What the data say about echo chambers (and where they diverge)

The literature uses "echo chamber" in different ways, and the choice of measure affects the answer observed. Cinelli et al. (2021) define an echo chamber as an environment in which the user's opinion, political leaning, or belief about a topic is reinforced by repeated interactions with peers or sources with similar tendencies and attitudes, and they operationalize it through two observable quantities, homophily in interactions on a topic (users who interact with like-minded others) and bias in information diffusion from like-minded peers, which must be quantified independently. They analyzed more than 100 million pieces of content on controversial topics (such as gun control, vaccination, and abortion) on four platforms, Gab, Facebook, Reddit, and Twitter, with one dataset per platform: abortion (Twitter), vaccines (Facebook), politics (Reddit), and the whole of Gab. Users' leaning, a value from −1 to +1 on the topic, was inferred from likes on labeled pages (on Facebook, for example, anti-vaccine counts as +1 and pro-vaccine as −1), from links to outlets classified by Media Bias/Fact Check (2,190 outlets), and from comments. Homophily was measured by the correlation between each user's leaning and the average leaning of his or her neighbors in the interaction network, and diffusion bias by the average leaning of the users reached from users of each leaning, in a simulation with an epidemic model (SIR, taken up again in Chapter 11). The authors contrast platforms on which users can tune their feed (Reddit) with those that do not offer this option (Facebook and Twitter), and they attribute to Gab's low moderation what they describe as widespread hate speech. On Facebook (vaccines) and Twitter (abortion), users' leaning correlated strongly with that of their neighbors, and information reached mainly those with a similar leaning; on Reddit and Gab, users formed a single community, leaning left (Reddit) or right (Gab). Because the differences between platforms might reflect the topics, the authors compared Facebook and Reddit on a topic both platforms shared, news consumption, and segregation was greater on Facebook. Table 6.1 summarizes the material and results of the three studies that follow.

Nyhan et al. (2023) measure something else: what reaches the feed of the entire active adult Facebook population in the United States. Only 20.6% of users receive more than 75% of their content from like-minded sources, and only 32.2% get a quarter or more from cross-cutting sources; according to the authors, the results do not bear out the worst fears about echo chambers. In the experiment, exposure to cross-cutting sources rose from 20.7% to 27.9%. Those whose exposure was reduced engaged more, per exposure, with the

like-minded content they still saw, and the authors state that the changes to the algorithm that they tested do not cancel out users' propensity to look for congenial information and engage with it. For the null effect they cite, among other explanations, that congenial political content is a fraction of what people see, that Facebook is a fraction of the information diet, and that the design does not capture effects of prior use or effects accumulated over years. They also note that reducing exposure might lead people to seek that information elsewhere on the internet.

Bail et al. (2018), in turn, ran the experiment in the opposite direction: they increased exposure to opposing views. A survey firm recruited self-identified Republicans and Democrats who visited Twitter at least three times a week; 1,652 answered a questionnaire in mid-October 2017 (901 Democrats and 751 Republicans), with 10 statements about public policy with which they were to agree or disagree on a seven-point scale that measures liberal–conservative position. One week later, the researchers randomly selected those who would be offered US$11 to follow, for one month, a bot, an automated Twitter account, that would retweet 24 messages a day, without the participant knowing its content. The bots, one liberal and one conservative, retweeted messages drawn at random from a list of 4,176 political accounts (elected officials, opinion leaders, media outlets, and nonprofit organizations). Of the Democrats, 64.9% accepted the offer, as did 57.2% of the Republicans, and the same questionnaire was administered again about a month and a half after the first. Republicans who followed a liberal bot expressed more conservative views: the increase was 0.12 points on the scale by the most cautious estimate, which counts everyone assigned to treatment, whether they followed the bot or not, and 0.60 among the fully compliant, those who correctly answered all the weekly questions about the content of the bot's tweets. Among Democrats, the effects were small, in the liberal direction, and not significant. The authors caution that they did not identify the mechanism, that the participants were paid to read messages from people who think differently, and that the findings should not be read as evidence that exposure to opposing views increases polarization in all contexts. Their second preregistered hypothesis is that people exposed to messages that conflict with their attitudes tend to counterargue them through motivated reasoning, which accentuates perceived differences and commitment to prior beliefs.

TABLE 6.1

Three studies on echo chambers: what they measured, what they observed, and what they do not allow us to say

Study	Material	What they measured	What they observed	Caveat
Cinelli et al. (2021)	More than 100 million pieces of content from Gab, Facebook, Reddit, and Twitter; leaning inferred from likes, links, and comments	Homophily in interactions on a topic and bias in diffusion from like-minded peers	Aggregation into homophilic groups dominates the dynamics; more segregation on Facebook (vaccines) and Twitter (abortion) than on Reddit and Gab, where users form a single community	Measures interaction and diffusion, not the effect on what users say or declare (addition); leaning is inferred
Nyhan et al. (2023)	2020 data on the entire active adult Facebook population in the U.S.; three-month experiment with 23,377 users	Share of content from like-minded and cross-cutting sources in the feed; eight preregistered attitude measures	Median of 50.4% (like-minded) and 14.7% (cross-cutting); exposure to like-minded sources from 53.7% to 36.2%, with no measurable effect on attitudes	Congenial political content is a fraction of what people see; Facebook is a fraction of the diet; the design does not capture effects of prior use or effects accumulated over years
Bail et al. (2018)	1,652 Twitter respondents (901 Democrats, 751 Republicans); US$11 to follow, for one month, a bot that retweeted 24 messages a day	Political opinions on a seven-point scale after the month	Treated Republicans: increase of 0.12 points (most cautious estimate) and 0.60 (fully compliant), in the conservative direction; Democrats: small effects, in the liberal direction, not significant	Mechanism not identified; participants paid to read messages from people who think differently; does not hold for all contexts

Note. The figures come from the articles. Cinelli et al. (2021) and Bail et al. (2018) were read in PMC XML, without pagination; the page numbers for Nyhan et al. (2023) are those of Nature. "(addition)" marks a caveat that is not the authors'.

The three studies seem to diverge, and in part they do, but they measure different things: whom people interact with on a topic (Cinelli et al., 2021), what reaches the whole feed, across all subjects (Nyhan et al., 2023), and what people declare when they are made to see the other side (Bail et al., 2018). It is possible, without contradiction, for a user to interact homophilically on a topic, receive a feed that is not very homogeneous overall, and move, on seeing the other side, toward where he or she already was. This is interpretation, not a finding of any of the three.

What they have in common matters more: none measures the consequences the audience arranges for what the speaker says. Exposure, interaction, and engagement tell us what arrives and whom one talks to, not who reinforces what. The closest datum is Nyhan et al.'s finding on engagement: those who saw less congenial content engaged more, per exposure, with what they still saw. One may read this as meaning that the propensity to seek the congenial is a function of a history of social consequences and not of what the feed offers, and one may also entertain the rival hypothesis that users simply like what they like. The study does not distinguish between the two.

DATA	There is strong homophily in interactions on specific topics on Facebook and Twitter (Cinelli et al., 2021); the median Facebook user's feed is half like-minded sources, only one fifth of users exceed 75%, and when congenial exposure was reduced by a third there was no measurable effect on attitudes (Nyhan et al., 2023); forced exposure to messages from the opposing camp moved Republicans in the conservative direction and did not significantly move Democrats (Bail et al., 2018).
INTERPRETATION	The three studies measure what arrives, whom people interact with, and, in two of them, what people declare after their exposure is changed. They do not measure the consequences the audience arranges for what the speaker says or how much of the correction reaches him or her, so they do not settle whether the bubble operates as a verbal community.
HYPOTHESIS	If the bubble operates as a verbal subcommunity, the effect on what is said depends on who reinforces and on how much of what comes from outside reaches the speaker, and not only on what he or she sees. No study in the chapter tests this.

6.3 “Virtual realities”: when verbal control replaces contact with the event

This section has no data: it is Guerin’s (1992) theoretical analysis and the extrapolation made here. Guerin describes three routes by which socially constructed knowledge becomes detached from the environment. In the first, social control of the tact passes to a group that does not reinforce accurate reporting of the environment. In the second, intraverbals of specific scope are reinforced by the group as if they were remembered tacts. In the third, generalized functional consequences become too generalized and maintain the tact indiscriminately. Guerin illustrates the second route with statements such as “nuclear reactors are unsafe”: most people have no direct experience with these events, and the statements are maintained by social consequences. Thus, under socioeconomic conditions that favor the use of fossil fuels (a favorable political context, low international prices, geopolitical conflicts, etc.), there will be a greater likelihood of reinforcement for such statements regardless of any additional factual information about the safety of nuclear reactors. In 1994 he returns to the example when discussing attitudes about events the person has never experienced, which are therefore intraverbals.

One of Guerin’s points deserves to be stated plainly, because it rules out a mistaken reading: even when what is said is counterfactual, the contingencies involved are real, albeit social (1992). Someone who claims there is a pink elephant in the room is under the control of social contingencies, not of the elephant. The expression “virtual realities,” from this book’s table of contents, is a shorthand for these routes, and none of the verified passages from Guerin uses it; “virtual” does not mean digital here, and the article predates social media.

Two passages matter for the digital environment. In the first, Guerin writes that mass media can sustain counterfactual social representations even without a social group, and that reinforcing verbal communities could potentially exist solely in the media rather than in person. In the second, he defines the first limit of socially constructed knowledge: the extent to which behavior is maintained by nonsocial stimulus control. The consequence drawn from the first limit is an extrapolation: where the speaker has direct contact with the event, the report comes under its control, and the gap between what the group reinforces and what the event shows tends to become visible. Guerin also discusses the limits through the distinction between rule-governed and contingency-shaped behavior (1992); Chapter 3 deals with it. The second limit is the subject of the next section.

6.4 Countercontrol and avoidance of countercontrol

Guerin (1992) points to a limit on the extent to which behavior can be maintained by a verbal community: strong pressure to conform generally gives rise to countercontrol when the contingencies it specifies are counterfactual, unless it is confined to a small, cohesive group, which works only when the benefits of group membership outweigh those of reporting accurately. The passage calls for a term and a question. The term is countercontrol. The question is what a group can do to keep the conditions for countercontrol from arising.

CONCEPT IN USE

Countercontrol (Delprato, 2002)

Delprato (2002) defines countercontrol as episodes of behavior made up of aversive controlling conditions mediated by other people and of escape or avoidance responses that fail to reinforce the behavior of the controller. It is not a basic principle of behavior: it is always escape or avoidance, and it is unique only in that the behaver, confronted with aversive social or interpersonal stimulation, responds so as to oppose the control instead of reinforcing it by yielding. Its forms include moving out of the controller's range, attacking, and passively resisting and, among many others, absenteeism, criticism, sabotage, noncompliance with requests, inattention, and quitting. It occurs at two levels: escape from or avoidance of specific short-term consequences, or of generalized, long-standing coercive contingencies. Delprato links effective countercontrol to conditions of control: control must be conspicuous and balanced with countercontrol, its sources must be identified, and their behavior must be accessible to those controlled; disguised control and the delegation of control should be avoided, because they make the controller's behavior inaccessible.

Two points in the text matter for what follows. The first is that countercontrol is a response to aversive control. Delprato (2002) writes that people prize freedom because it refers to nonaversive methods of control, which for that reason do not occasion countercontrol, and that those whose behavior is positively reinforced report feeling free because the variables controlling them may go unnoticed. For the bubble, if the hypothesis of Section 6.1 is correct, the consequence is that the control exerted on members, made up largely of approval, tends not to occasion countercontrol within the group, as in Guerin's small, cohesive community. The countercontrol that matters comes from outside or from dissenters: an agency's correction, criticism from those who

disagree, a platform's moderation, a member's departure. The second point is the list of conditions cited by the author and examined below, which the chapter uses as a yardstick.

There is also a question of direction. In Delprato, countercontrol is the response of the one who is controlled. Avoidance of countercontrol, as this chapter uses the term, is the other side of the relation: what the speaker, or the community that reinforces the speaker, does that reduces the chance that this response will occur, or its effectiveness. It is an extension proposed here, and the box below says so plainly.

CONCEPT IN USE

Avoidance of countercontrol (proposed here, based on Delprato, 2002)

A repertoire of the speaker, or of the community that reinforces the speaker, maintained because it reduces the probability or the effectiveness of countercontrol: correction, criticism, denunciation, moderation, departure. The term is not Delprato's, nor that of any other source read. Calling a repertoire avoidance requires showing the contingency: that correction is aversive to the speaker or the community, that the repertoire precedes or prevents it, and that it is maintained for that reason and not merely by the audience's approval (Section 6.8). Four repertoires are discussed: form, source, place, and people.

Form. Guerin (1994) writes that, when certain attitude statements carry negative consequences, autoclitics offer several ways of avoiding them. Reporting attitudes is reinforced when supporting beliefs are presented, but only if the verbal community agrees with them; in front of an audience that does not hold them, the belief would be omitted. Chapter 2 (Section 2.8) argued that a statement that shifts from factual to opinion form ceases to be checkable, and Guerin shows that attitude statements can function as mands. The reading as avoidance is that form strips the control of its conspicuousness: the statement that acts on the listener comes out as opinion. Guerin's datum, however, has the audience as the variable, which suggests that the protection afforded by form is more necessary at the fringes than inside the bubble. To a die-hard adherent of the group, the speaker says "vaccines don't work," whereas to people at the edge of the group he or she says "I don't believe vaccines work."

Source. Evangelista and Bruno (2019) conducted participant observation in at least three political WhatsApp groups and two Telegram groups, up to the day of the 2018 election and in the preceding weeks. In the groups clearly identified with Bolsonaro, they observed the repeated claim that the mainstream media, controlled by communists and leftists, should not be believed, and that only a few YouTubers, journalists, bloggers, and politicians deserved

trust. After the election, mainstream media content critical of the new government became less common in the groups, even when it came from commentators regarded as right-wing, whereas before, material critical of the PT (Workers' Party) had been shared. For the authors, the strategy seems to be to legitimize the groups as more trustworthy sources than the mainstream media. During the pandemic, Bonow Soares et al. (2021), who analyzed the most-shared messages from more than 500 public political WhatsApp groups in March and April 2020 (802 messages with disinformation about Covid-19), state that political argumentation served to cast every discourse that discredited Bolsonaro as a creation of the opposition. Ricard and Medeiros (2020) report, according to the joint parliamentary commission of inquiry (CPMI) of the National Congress that has been investigating disinformation since 2019, that part of the disinformation attacked governors, the media, and health experts in order to delegitimize those who supported social isolation. The change within the same type of source, shared or not depending on whom it criticizes, is the closest datum to control by the audience: what is shared varies with the target of the criticism, not with the history of the source.

Place. In Agência Lupa's 2026 report (Lupa is a Brazilian fact-checking agency), the electoral fact-checks are those of viral or potentially viral content without identified authorship, in the electoral cycles from 2018 to 2024 (286 in all). Facebook appeared in 76.5% of the 2018 fact-checks, WhatsApp in 92.4% of those from 2022 and in 100% of the electoral fact-checks analyzed in 2024 (Agência Lupa, 2026). The agency reads the datum as a migration of misleading content, as open networks expanded moderation and came under greater public scrutiny, to messaging apps and semi-private environments, where authorship is harder to identify and monitoring faces greater limitations. Recuero et al. (2025), who analyzed 811 fact-checks by three Brazilian agencies on the 2018 and 2022 elections, read the shift from Facebook to WhatsApp and Twitter between 2018 and 2022 as an adaptation to more closed spaces, in which it is harder for counter-discourses to circulate, and Bonow Soares et al. (2021) write that WhatsApp is a key channel for spreading false information because it cannot be tracked or publicly debunked by fact-checkers. The denominators are small (51 fact-checks in 2018, 131 in 2022, and 54 in 2024, read from the chart), and Lupa itself warns that it maintains a WhatsApp tipline and takes part in Meta's fact-checking programs, so the results about platforms are a picture of the universe it monitors.

People. In the groups Evangelista and Bruno (2019) observed, critical comments led to the removal of their author and to the accusation of being an infiltrated petista (PT supporter); phone numbers of the accused circulated, and moderators were alerted to remove them. There was little discussion, and a

few members accounted for almost all the content sent or forwarded. The authors point to the invisibility of the actors who produce, monitor, distribute, or direct the content: most users do not know the authorship of what they receive and cannot verify it. Lupa writes that the absence of identifiable authorship of the attacks on candidates hinders accountability and favors the rapid spread of this content (Agência Lupa, 2026). The figures on the concentrated origin of the images come from Resende et al. (2019, as cited in Evangelista & Bruno, 2019). These researchers monitored public WhatsApp groups chosen by matching the invitation link against a dictionary of names of politicians, of parties, and of words associated with political extremism: 141 groups during the truckers' strike and 364 during the election. An automatic method identified the shared images that had already been debunked by fact-checking services, and to these were added 15 that Lupa had identified as disinformation. These 85 images with disinformation were shared eight times more than another 69,590, true or not reported to any independent agency; they made up 1% of the images and appeared in 44% of the groups monitored during the campaign, and a small number of groups seemed to originate a large share of them.

TABLE 6.2

Four repertoires of avoidance of countercontrol: the Delprato condition each one weakens and the sources that describe it

Repertoire	Condition for effective countercontrol that it weakens (Delprato, 2002)	Sources that describe it
1 Form	Conspicuous control: the statement that acts on the listener comes out as opinion or attitude	Guerin (1994)
2 Source	None: not on Delprato's list; an addition proposed here	Evangelista and Bruno (2019); Bonow Soares et al. (2021); Ricard and Medeiros (2020)
3 Place	Controller's behavior accessible to those controlled	Agência Lupa (2026); Recuero et al. (2025); Bonow Soares et al. (2021)
4 People	Identified sources; no delegation of control	Evangelista and Bruno (2019); Agência Lupa (2026)

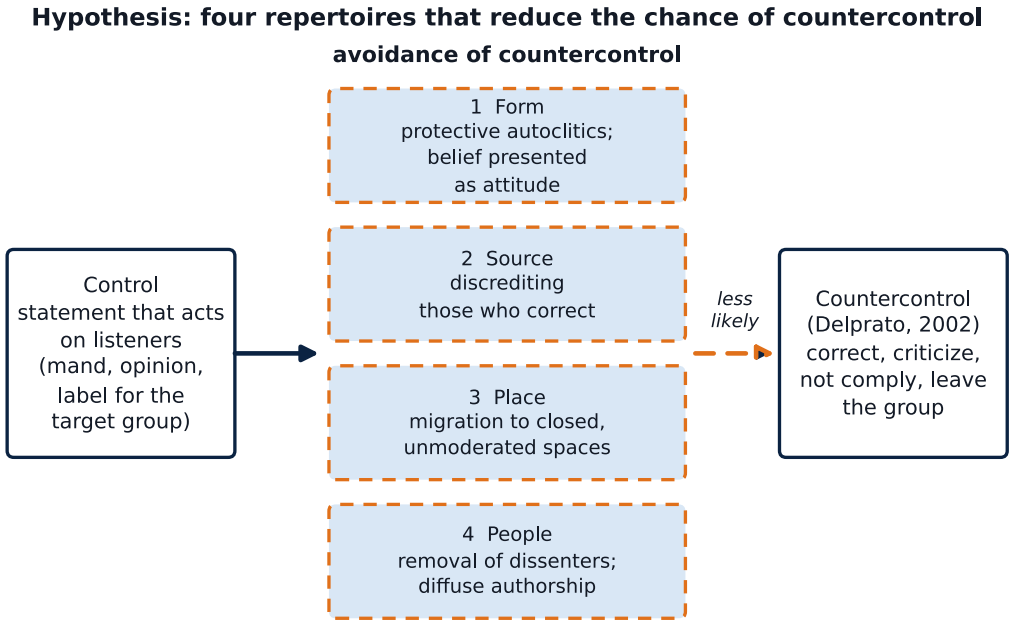
Note. Page numbers from the journals' PDFs and from Lupa's report. The classification as an avoidance repertoire is an interpretation proposed here; each source's caveats are in the text of the section.

The figure summarizes the hypothesis; it does not prove it. Repertoires 1, 3, and 4 weaken conditions on Delprato's list: form makes control less conspicuous, place makes the controller's behavior less accessible, and diffuse authorship makes the source less identifiable, while the removal of dissenters elim-

inates those who would exert countercontrol. Repertoire 2 is not on the list; it acts on the effectiveness of countercontrol, not on its conditions, by discrediting those who exert it, and that is why it is treated as an addition. That the four function as avoidance in the technical sense, maintained because they prevent correction, is the part that no source in the chapter tests: the sources describe the repertoires and, at most, the authors' own reading, without recording the correction and what the speaker or the group did afterward.

FIGURE 6.2

Control, countercontrol, and the avoidance repertoires: a hypothesis



Repertoires 1, 3, and 4 weaken conditions that Delprato (2002, p. 196) associates with effective countercontrol (conspicuous control, accessible controller behavior, identified source). Repertoire 2 is an addition proposed here.

Dashed: hypothesis proposed here. The term "avoidance of countercontrol" is not Delprato's.

Note. Original diagram by the author, without data. Dashed: hypothesis. The term "avoidance of countercontrol" is not Delprato's (2002). Repertoires 1, 3, and 4 weaken conditions he associates with effective countercontrol; repertoire 2 is an addition proposed here.

DATA	In political WhatsApp and Telegram groups observed in 2018, criticism led to the removal of its author, a few members sent almost all the content, and the mainstream media were described as untrustworthy (Evangelista & Bruno, 2019). The share of Lupa's fact-checks involving WhatsApp rose from 2018 to 2024 (Agência Lupa, 2026). Guerin (1994) describes autoclitics as avoiding negative consequences of attitude statements.
INTERPRETATION	The repertoires described are compatible with what Delprato (2002) associates with ineffective countercontrol: disguised control, unidentified sources, and inaccessible controller behavior. That they reduce the chance or the effectiveness of correction is a reading; that they are maintained only by the audience's approval is an equally compatible reading.
HYPOTHESIS	The repertoires of form, source, place, and people are maintained, at least in part, because they prevent correction. This predicts that the form of speech will change after fact-checks, that discrediting will accompany or precede correction, and that the removal of dissenters or a change of place will follow it. No source in the chapter records the correction and what came afterward. Although these repertoires are presented here in the context of disinformation, it is the authors' position that they are present in any social interaction.

6.5 The Brazilian case: WhatsApp, elections, and the pandemic

The chapter's Brazilian material is scant and comes from different designs. Table 6.3 brings it together before any interpretation.

TABLE 6.3

The chapter's Brazilian material: what each source contains, what it records, and what it does not allow us to say

Source	Material	What it records	What it does not allow us to say
Evangelista and Bruno (2019)	Participant observation in at least three political WhatsApp groups and two Telegram groups, up to the 2018 election and in the weeks before; data from Resende et al., cited by the authors	Removal of those who criticized; a few members sent almost all the content; mainstream media described as controlled by communists and leftists; invisible authorship	Few groups, no sampling described: does not measure prevalence; the authors say they cannot fully confirm the effective use of microtargeting
Bonow Soares et al. (2021)	Most-shared messages from more than 500 public political WhatsApp groups (Mar–Apr 2020): 802 with disinformation about Covid-19, according to the authors	41% conspiracy theories, 39% distortions, 20% fabricated information; among the topics (not mutually exclusive), 36% governors and mayors with hidden interests against Bolsonaro and 21% the media	Only public groups, only text, about one third of the messages shared at least 20 times a day; many groups are of Bolsonaro supporters
Ricard and Medeiros (2020)	Commentary, without original data; testimony and CPMI	Part of the disinformation attacked governors, the media, and health experts to delegitimize those who supported social isolation	Not an empirical study; the commission had not yet released evidence of direct coordination by the so-called Office of Hate (Gabinete do Ódio) or by members of the government
Agência Lupa (2026)	286 electoral fact-checks (viral content without identified authorship), from 2018 to 2024	Facebook in 76.5% (2018); WhatsApp in 92.4% (2022) and 100% (2024); 73.1% about voting machines, vote counting, the TSE (Superior Electoral Court), or results; 50% involving public institutions	A picture of the universe Lupa monitors (WhatsApp tipline, Meta programs); categories not mutually exclusive
Recuero et al. (2025)	811 fact-checks by three agencies, 2018 and 2022 elections	Change of platform read as adaptation to more closed spaces; both campaigns appear as producers and victims	A study of fact-checks, not of groups; the authors' reading

Note. Page numbers for Evangelista and Bruno (2019), Bonow Soares et al. (2021), Ricard and Medeiros (2020), Agência Lupa (2026), and Recuero et al. (2025) are those of the journals' PDFs and of the report. Lupa's denominators, from 51 to 131 fact-checks per year, are read from the chart on p. 80.

Two things can be read in the table. The first is what recurs across sources with different designs: the disqualification of those outside the group, of those who correct, or of those who support an opposing policy (Evangelista and Bruno; Bonow Soares et al.; Ricard and Medeiros), and authorship that is hard to identify (Evangelista and Bruno; Agência Lupa). During the pandemic, Bonow Soares et al. (2021) record that disinformation in the monitored groups was often linked to conspiracy theories, such as alleged plots by leftists, the STF (Brazil's Supreme Federal Court), and the media against Bolsonaro. Two coders classified the 802 messages, with a third to break ties: 41% were conspiracy theories, 39% distortions, and 20% fabricated information, in mutually exclusive categories and with an intercoder reliability (Krippendorff's alpha, a measure of agreement) of 0.601, which the authors call moderate to substantial. Among the topics, which could overlap, 36% dealt with governors and mayors with hidden interests against Bolsonaro and 21% with the media. In the election, half of Lupa's fact-checks involve public institutions, the judiciary, the Armed Forces, or security agencies, often presented as participants in alleged conspiracies or irregularities (Agência Lupa, 2026).

There is a hypothesis into which these numbers fit, and it is proposed here: if the institutions that tally and verify the election (the electoral court, the judiciary, the press) are the sources of countercontrol for claims about electoral events, presenting them in advance as participants in conspiracies is a way of weakening the countercontrol that would come from them, the source repertoire of Section 6.4. The rival reading is that the content simply deals with what there is to deal with in an election: the subject of voting machines is the subject of the election, and the institutions are its object. The material does not separate the two readings.

The second is balance. Almost all the chapter's Brazilian material deals with groups and campaigns of one political camp, that of Bolsonaro's supporters, and this follows from the design of the sources (Bonow Soares et al., 2021; Evangelista and Bruno, 2019), not from a choice of camp, since the same method applies to all. Recuero et al. (2025) note that the literature they cite points to more disinformation produced by Bolsonaro's supporters in both elections, but that both campaigns appear as producers and victims of disinformation. Evangelista and Bruno (2019) state that, in Resende et al.'s sample, there is no evidence that the left-wing coalition used the same tactics, although there are reports that the campaign of Fernando Haddad (the PT candidate in 2018) also

used WhatsApp. The two records do not settle whether the phenomenon is asymmetrical: they say that the question, for the avoidance repertoires, has not been asked with a comparable sample. The chapter's hypothesis is one of mechanism: the repertoires are defined by their relation to correction, not by party, and should appear wherever there is a cohesive group under the control of an audience that reinforces it. Absence of data is not absence of the phenomenon, and the chapter does not claim symmetry; it claims that the data do not allow us to say.

DATA	In political messaging groups observed in 2018 and in public WhatsApp groups monitored in 2020, the sources describe removal of critics, few producers, invisible authorship, distrust of the mainstream media and, during the pandemic, conspiracy theories involving governors, the media, leftists, and the STF (Evangelista & Bruno, 2019; Bonow Soares et al., 2021). Half of Lupa's electoral fact-checks involve public institutions, often presented as participants in alleged conspiracies (Agência Lupa, 2026).
INTERPRETATION	What is described is compatible with the four repertoires of Section 6.4, and the classification is proposed here. The sources deal, for the most part, with one political camp, by design.
HYPOTHESIS	The repertoires will appear, with the same function, in groups of other camps when the same sampling rule is applied to them (see the protocol in Chapter 2, Section 2.9). There are no data on this in the chapter.

6.6 Who likes, who replies: an exploratory analysis of two political videos on Instagram

In October 2026, during the campaign for the runoff of the presidential election, a series of short videos about politics was published on the Instagram profile of one of the authors (@marcioborgesmoreira). The profile had 11,550 followers and lived on science-communication carousels about behavior analysis, generally seen by a few hundred accounts. Two videos in the series were outliers: "Eram só vovós com Bíblia?" ("Were they just grannies with Bibles?"), about the January 8, 2023, attacks in Brasília, had 1.29 million views, and "Quanto custava o litro da gasolina em cada ano?" ("How much did a liter of gasoline cost each year?"), 735,000. The author expected attacks. The topic, the polarization of the campaign, and an audience a hundred times larger than the follower list seemed to bring together the conditions of a negative audience, in Skinner's sense, taken up in Section 10.1. The attacks did not come: in

Direct messages, nothing; in the comments, as far as one could see, almost nothing but approval. The impression was that the videos had gone to a million people in the author's bubble, that of Lula's camp.

This section examines that impression with the profile's own data, and the case matters to the chapter for a precise reason. The studies in Section 6.2 measure what reaches the feed and whom people interact with; none measures the consequences the audience arranges for what is said, or what happens to those who disagree. On a profile, the person who posts has access to part of this: how many accounts saw it, how many were followers, what each person wrote, how many likes each comment received, and what replies it provoked. The case does not test the chapter's hypothesis, because it is one profile and one episode, but it shows what is gained by describing the bubble through its consequences, and where data are still lacking.

There is a caveat about origin, and it comes before the numbers: the videos belong to one of the authors, who had a position in the campaign. This does not change the book's rules of reading (data, interpretation, and hypothesis kept separate; the same criterion for all camps), but readers should take it into account when judging the classification of the comments, which is interpretation.

The material and how it was obtained

The material covers the 74 posts published on the profile between September 21 and October 21, 2026, and was read on October 21, between 8 and 10 p.m. (Brasília time). It came from three sources. The Instagram Graph API, the interface Meta offers to professional accounts, provided the date, type, caption, likes, and comment count of each post and, for the two videos, the text of all comments and of the replies to them. The access token available did not have permission to read metrics; for that reason, views, reach, the split between followers and nonfollowers, shares, saves, average watch time, and the audience by age and gender were read on each post's page in Meta Business Suite, Meta's own dashboard, and the followers' profile on the audience page of the same dashboard.

Two platform measures need to be explained before they are used. Reach is the number of distinct accounts that saw the post; a view is each time it starts to be displayed, which is why views exceed reach (in the January 8 video, 1.37 per account). The platform also splits views and reach between followers and nonfollowers of the profile, and the two splits do not add up: the page of the January 8 video attributes 224,100 views to followers, more than the two accounts that published it (with 11,550 and 5,198 followers) would produce even

with repeat views, while the account page records 5,202 followers reached in the whole month. Since Meta does not explain the difference, the split is used here only for its direction, which the two measures agree in showing.

The comments came back incomplete, and the gap is itself a datum. Of the 2,529 the platform counted on the January 8 video, the API returned 2,076 (1,901 top-level comments and 175 replies), written by 1,778 accounts; of the 962 on the gasoline video, 876 (637 and 239), from 705 accounts. Missing are 453 (18%) and 86 (9%). None of those returned was flagged as hidden, and the API does not say what is missing: comments deleted by their authors, restricted accounts, and comments held by the platform's automatic filters are possible explanations, which the app's dashboard would allow one to check and which were not checked. The comments, with usernames, were kept only on the author's server; no names appear here, and hostile remarks are paraphrased.

Each comment received three codes, following the rules in Box 6.1: its position with respect to the video, the political camp it declares, and, if present, hostility and its target. The reading was done by a single coder, the artificial intelligence assistant used in preparing this book, comment by comment (2,457 of them); another 495, consisting only of a mention of another account, only a sticker, only applause or laughter emojis, or only a candidate's slogan, were classified by rule. There was no second, independent coding, and intercoder agreement was not measured; the numbers in Table 6.5 hold subject to this caveat.

BOX 6.1

Coding rules for the comments

Position with respect to the video (one per comment):

- **Agreement with the content:** agrees with what the video shows or argues, praises, thanks, adds an argument or fact in the same direction, or mocks the version the video contests.
- **Campaign expression:** vote, slogan, hashtag, or request for votes, with no reference to the video's content.
- **Contestation:** disagrees, corrects, counterargues, relativizes, or counters with another case ("what about the MST?", the Landless Workers' Movement), whatever the camp of the person contesting.
- **Attack on the author:** insult, personal disqualification, or accusation of bad faith directed at the person who made the video.
- **No position:** only a mention of another account, a sticker, an ambiguous emoji, a neutral question, a report of a local price without contestation, an unrelated subject.

Declared camp: Lula and the left, Bolsonaro and the right, or none; only when the comment says so (vote, slogan, support for or attack on a candidate by name, a label for the other camp, such as "gado" ["cattle"] or "petralha" [a slur for PT supporters]).

Hostility and target: insult, derogatory label for a person or group, dehumanization, wish for harm or for punishment beyond what the courts decided, mockery of a personal attribute (age, appearance). Targets: participants in the January 8 attacks; right-wing voters or activists; right-wing politicians; Lula, the PT, the left, or their voters; the author; another commenter; other. Terms that describe acts ("vandals," "coup plotters," "criminals") and judgments about ideologies ("Bolsonarism is a cult") do not count. "Hate speech" is not a code: in this book, it is the label of an external authority (Chapter 5).

Irony: recognized only when marked (laughter, quotation marks, "contains irony") or when the literal reading contradicts the comment itself; otherwise, the literal reading holds.

Disguised spelling: a letter replaced by a number, @, or asterisk, or a word broken up by periods or spaces ("c@gona," "L U L A"). An automatic rule finds the candidates; reading decides.

What the profile's numbers show

The profile had a routine before the series, and it is against that routine that the two videos are measured. For the 45 science carousels in the period, the median was 464 views and 202 accounts reached, and only 3.7% of the views came from people who did not follow the profile: these were posts seen by the base. The political series began on the night of October 14 and totaled 24 Reels in seven days. The 14 data Reels, all in the format "E quem estava na Presidência" ("And who was President": an annual series from an official source, with a photo of each president), had a median of 3,300 views, 86.8% of them from nonfollowers; the other 8 political Reels, 2,400 and 76.8%. The two outlier videos totaled 2.02 million views, 93% of everything the 74 posts had in the month (Table 6.4 and Figures 6.3 and 6.5).

TABLE 6.4

The two videos and the profile's routine: medians by type of post, 9/21 to 10/21/2026

Measure	Science carousels (45)	Data Reels (14)	Other political Reels (8)	January 8 video	Gasoline video
Views	464	3.3K	2.4K	1,293,276	735.5K
Accounts reached	202	2.6K	2.0K	947,057	536.3K
Views from nonfollowers	3.7%	86.8%	76.8%	82.5%	93.9%
Likes	17	61	63	77,673	17,015
Comments	0	1	1.5	2,529	962
Shares	1	12.5	14	23,232	14.7K

Measure	Science carousels (45)	Data Reels (14)	Other political Reels (8)	January 8 video	Gasoline video
Saves	6	9	9	8,864	2.1K
Average watch time	not applicable	16 s	17 s	26 s (of 59 s)	25 s (of 49 s)
Men among accounts reached	49.7%	79.7%	68.4%	54.0%	84.9%
Followers gained	0	1.5	3	144	735

Note. Medians in the first three columns. Reels published from 10/14 to 10/21/2026; the carousel column covers the whole month, and the nonfollower share is based on 44 posts, because the page of one of them did not show it. For posts not cross-posted to Facebook, the platform gives reach rounded ("2.6K"). The January 8 video was published in collaboration with another account, with 5,198 followers. Sources: Instagram Graph API and Meta Business Suite, read on 10/21/2026; full table, by post, in `verificacao/caso-instagram-2026/posts_metricas.tsv`.

The change of scale shows up in full in the account's daily reach (Figure 6.4). Until October 13, the profile reached 200 to 800 accounts a day; the only spike in the period was a five-day paid ad, from September 30 to October 4, for a post from before the series (R$139.92, 23,502 views). On October 15, the first full day of the January 8 video, it reached 563,637 accounts. Neither video was boosted. The January 8 video passed 500,000 views about 19 hours after it was published and 1 million in two and a half days; the gasoline video reached 500,000 in about 39 hours and 700,000 in four days. In the 28 days from September 23 to October 20, the account reached 1.5 million accounts, of which 5,202 were followers (0.3%) and 1,512,615 were not (99.7%); 97% of the interactions came from nonfollowers.

FIGURE 6.3

The two outlier videos and the series they belonged to

"Were they just grannies with Bibles?"
Oct 14, 11:29 p.m., collaboration



views **1,293,276**	likes **77,673**
accounts reached **947,057**	comments (read) **2,529 (2,076)**
from nonfollowers **82.5% of views**	shares **23,232**
average watch time **26 s of 59 s**	saves **8,864**

"How much did a liter of gasoline cost?"
Oct 17, 11 a.m.



views **735.5K**	likes **17,015**
accounts reached **536.3K**	comments (read) **962 (876)**
from nonfollowers **93.9% of views**	shares **14.7K**
average watch time **25 s of 49 s**	saves **2.1K**

The same series, on the same days: eight data Reels and the views of each one

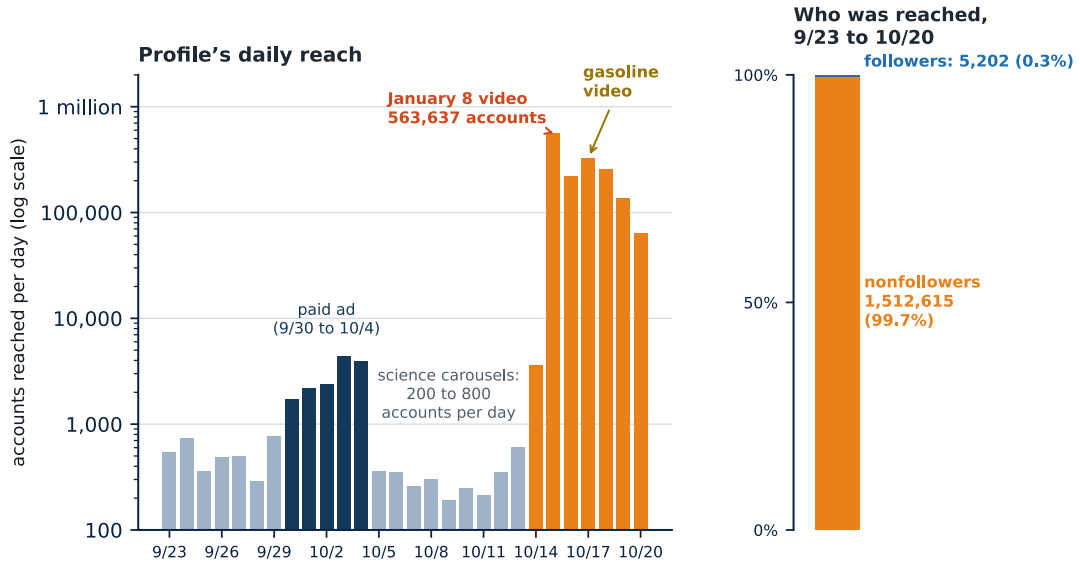


6.9K · 4.3K · 8.1K · 1.9K · 4.5K · 5.3K · 2.0K · 5.0K

Note. Covers of the author's posts, obtained through the Graph API. On the January 8 cover, the top half was blurred: it shows an identifiable person, who is not a public figure, in a video frame bearing a newspaper's logo. Official photos on the other covers: Ricardo Stuckert/PR, Roberto Stuckert Filho/PR, Beto Barata/PR, and Alan Santos/PR (CC BY), credited in the posts' captions. Numbers from Meta Business Suite and the Graph API, read on 10/21/2026.

FIGURE 6.4

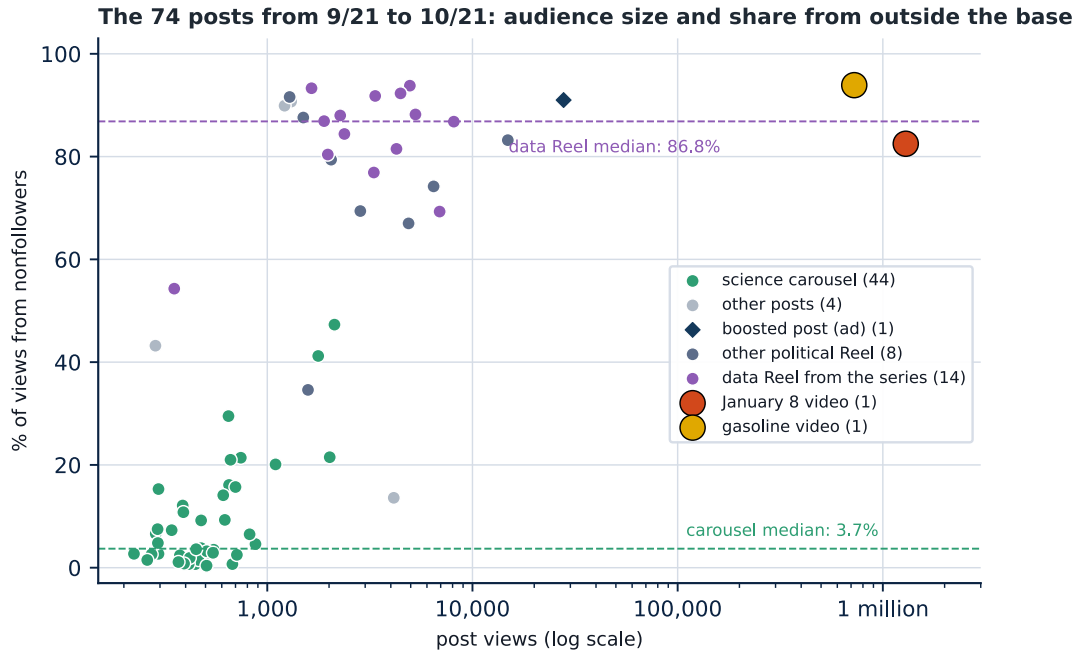
The profile's daily reach and the share of followers and nonfollowers, 9/23 to 10/20/2026



Note. Days counted in Pacific time, as in Meta's dashboard: the January 8 video, published at 11:29 p.m. on 10/14 Brasília time, falls at the end of 10/14 in that time zone. Source: Meta Business Suite, Instagram account overview, read on 10/21/2026.

FIGURE 6.5

The 74 posts in the period: audience size and the share coming from outside the base

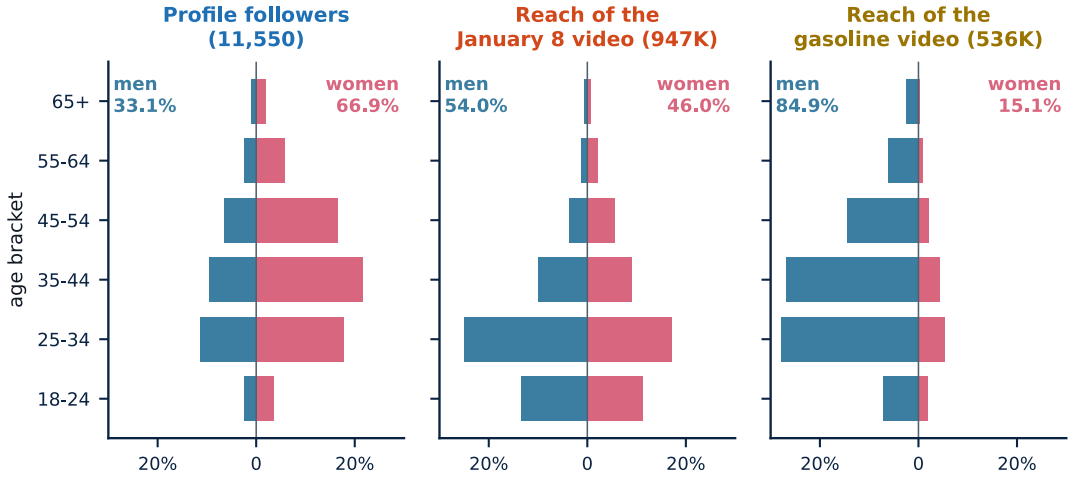


Note. Each point is a post; the nonfollower share is that of views, as the platform splits it. "Other posts": three institutional Reels (two about books from the Institute and one about Rádio iW4) and an image about voting. Source: Meta Business Suite, each post's page, read on 10/21/2026.

Those reached were not the base in another sense either, that of composition (Figure 6.6). The profile's followers are 66.9% women, and 83% of them are between 25 and 54 years old. Among the 947,000 accounts reached by the January 8 video, 54% belonged to men, and 67% were between 18 and 34; for the gasoline video, 84.9% belonged to men. Even compared with the audience the science carousels reached, which had a median of 49.7% men, the audience of the data Reels (79.7%) and that of the gasoline video were a different public.

FIGURE 6.6

Who follows the profile and whom the two videos reached, by age and gender



Note. Percentages of the total of each audience. Followers: audience page in Meta Business Suite; reach: each post's page. The platform infers age and gender from account registration. Read on 10/21/2026.

Two other measures tell what this audience did besides watching. It shared and liked a great deal, but hardly started following: the January 8 video gave the profile 144 new followers, one for every 6,600 accounts reached, and the gasoline video, 735; in the month, there were about a thousand new followers and 404 unfollows. And it did not write to the profile: the dashboard records 36 conversations started in Direct in 28 days, 16 of them during the seven days of the series, and 12 new contacts, nothing that keeps pace with a reach of 1.5 million accounts. The content of the messages was not read.

What the comments said

The two videos drew comments of very different composition (Table 6.5 and Figure 6.7). On the January 8 video, 52.1% expressed agreement with the content, 19.3% were Lula campaign comments, and 26.0% took no position; contestation totaled 2.2% (45 comments), half of it from people declaring Bolsonaro's camp, and only one comment attacked the author. Comments declaring

Bolsonaro's camp numbered 32 (1.5%), against 536 (25.8%) from Lula's camp. Agreement took recurring forms: mocking the "grannies" version ("tadinhos" ["poor things"], "só tinham santos" ["all they had were saints"]), repeating the line of the elderly woman shown in the video, noting that the participants themselves had filmed the images, drawing comparisons with left-wing protests, and asking that there be no amnesty. Campaign comments came as numbers, slogans, and hashtags against the opponent; a single account pasted the same voting message 30 times.

On the gasoline video, a third of the comments (35.7%) contested the video, and the contestation came from three sides. From those declaring Lula's camp came 117 comments saying the video was lying because, under the Bolsonaro government, a liter had reached 9, 10, or 12 reais: peak prices, from one gas station and one city, against an annual national average adjusted for inflation, which the cover and the caption announced as "em reais de hoje" ("in today's reais"). From Bolsonaro's camp came 118, almost all of them recalling the pandemic, the war in Ukraine, and the 2022 tax cut, which the video did not mention. And 78, with no declared camp, contested it with the nominal prices of the time ("em 2006 era 2,50" ["in 2006 it was 2.50"]), often copying a search engine's summary. Comments from Bolsonaro's camp were 13.9% of the total, nine times the proportion on the other video.

TABLE 6.5

The comments on the two videos by category, and the likes the audience gave to each

Category	January 8: comments	January 8: likes	Gasoline: comments	Gasoline: likes
Agreement with the content	1,081 (52.1%)	30,321 (91.0%)	202 (23.1%)	711 (24.3%)
Lula campaign	400 (19.3%)	820 (2.5%)	138 (15.8%)	414 (14.1%)
No position	539 (26.0%)	2,062 (6.2%)	217 (24.8%)	1,008 (34.4%)
Contestation from Lula's camp	1 (0.0%)	0	117 (13.4%)	565 (19.3%)
Contestation with no declared camp	22 (1.1%)	4 (0.0%)	78 (8.9%)	183 (6.2%)
Contestation from Bolsonaro's camp	22 (1.1%)	103 (0.3%)	118 (13.5%)	49 (1.7%)
Bolsonaro campaign	10 (0.5%)	3 (0.0%)	4 (0.5%)	0
Attack on the author	1 (0.0%)	0	2 (0.2%)	0
Total	2,076	33,313	876	2,930
With hostility, any target	265 (12.8%)	941 (2.8%)	110 (12.6%)	168 (5.7%)

Note. Categories from Box 6.1, with position and camp combined. Comments and replies returned by the Graph API (2,076 of 2,529 and 876 of 962); likes given to these comments up to the evening of 10/21/2026. On the gasoline video, the "no position" category concentrates likes because of a sticker that received 849. Coding by a single coder, with no measure of agreement. Source: `verificacao/caso-instagram-2026/comentarios_categorias.tsv`.

Hostility was present on both videos, in similar proportions (12.8% and 12.6% of comments), and almost all of it pointed outward. On the January 8 video, the targets were the participants in the riots (117 comments: insults to the elderly woman in the video, mockery of her age, calls for harsher sentences than those handed down, some death wishes), right-wing politicians (75), right-wing voters (31, with "gado" ["cattle"] and "bolsominions"), other commenters (24), Lula, the PT, or the left (5), and the author (1). On the gasoline video, other commenters (39, almost all in replies), right-wing politicians (32), Lula, the PT, or the left (23), right-wing voters (13), and the author (3). Across the two videos, four of 2,952 comments (0.14%) were hostile toward the person who published them; one of them, from someone who declared Lula's camp, accused the video of lying by showing the price under the Bolsonaro government as lower than he remembered it. Those who arrived with hostility generally directed it at the group the video criticized, not at the person who published it.

Who reinforces what: likes and replies

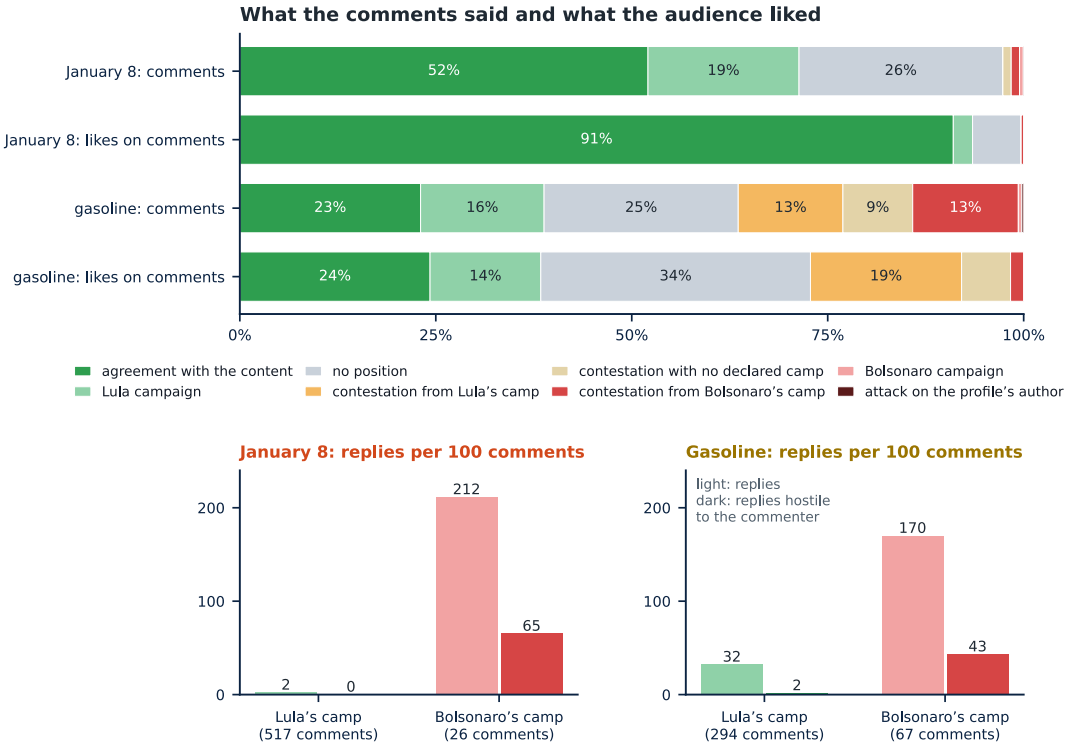
The audience does not only comment: it likes comments and replies to them, and these are the consequences it arranges, visible and countable, for what each person wrote. On the January 8 video, comments expressing agreement received 91% of all likes given to comments. The four most-liked, all expressing agreement, had 12,623, 8,608, 3,999, and 1,266 likes (the most-liked said "a melhor parte é que eles mesmos gravaram" ["the best part is that they filmed it themselves"]); since the platform orders comments by relevance, this was what one saw first on opening them, and my impression of "almost nothing but approval" probably came from this showcase. On the gasoline video, the two groups of partisan contestation were the same size, 117 and 118 comments, but those from Lula's camp received 565 likes and those from Bolsonaro's camp, 49. The audience reinforced contestation according to the camp of the person contesting, not according to the quality of the argument: recalling the pandemic is no less pertinent than recalling a peak price. Correction within one's own camp was also reinforced: a comment saying that under the Dilma government gasoline had never reached 4 reais got 169 likes, the reply explaining the inflation adjustment got 284, and the person who had commented replied "entendi" ("got it").

The replies show the other side of the same contingency (Figure 6.7). On the January 8 video, the 517 top-level comments from Lula's camp received, together, 12 replies, one of them hostile; the 26 from Bolsonaro's camp received 55, of which 17 were hostile toward the commenter. Per hundred comments, that is 2 replies and less than one hostile reply in one camp, and 212 replies and 65 hostile ones in the other. On the gasoline video, where discussion was greater for everyone, the difference persists: 32 replies and 2 hostile ones per hundred comments from Lula's camp, 170 and 43 per hundred from Bolsonaro's camp.

Two threads show how this happens. On the January 8 video, a commenter wrote that the participants "aprenderam com o MST" ("learned it from the MST"), with two laughing emojis. The comment got 94 likes and 47 replies: some argued (the MST produces food, donated food during the pandemic), and 15 were hostile toward her, including calling her ignorant and telling her to shut up, saying that her husband was the one who should give opinions, commenting on her skin color and her husband's profession and, in one of them, a veiled warning about the photos of her daughters that she posted. On the gasoline video, a commenter wrote that in his city a liter cost 5 reais under the Bolsonaro government and 7.80 now. He received 35 replies, 18 of them hostile, almost all mockery; the most-liked, with 77 likes, asked whether there were militias running the gas stations in his city. In a third episode, the comment "Flávio Bolsonaro é o nosso futuro presi..diário" (659 likes; a pun on presidente, "president," and presidiário, "convict") was read by some as support for Flávio and received replies such as "o seu, essa maldição" ("yours, you curse"), until someone pointed out that they had not read it properly: the audience was responding to what looked like dissent before reading it through.

FIGURE 6.7

What the comments said, what the audience liked, and what happened to those who disagreed



Note. Comments and replies returned by the Graph API; likes up to the evening of 10/21/2026. Below, only top-level comments with a declared camp and the replies they received; a hostile reply is one with hostility directed at the commenter (Box 6.1). Coding by a single coder. Source: `verificacao/caso-instagram-2026/comentarios_categorias.tsv` and `respostas_por_campo.tsv`.

Form, source, and people: the repertoires of Section 6.4 in the comment box

The four repertoires of Section 6.4 were described, in the chapter's Brazilian material, in messaging groups of a single political camp. Three of them appear in the comment box of these videos, in an audience aligned with the opposite camp, and the description that follows is topographical: it says what form they had and what consequences followed, not that the repertoire is maintained because it prevents correction.

Form. Twenty-eight comments (0.9%) disguised their spelling: "c@gona," "vaga@bund0s," "m4t4e," "ban.did0 bom e ba.ndid0 m0.rt0" ("a good criminal is a dead criminal"), "presi..diário," "L U L A." The disguise fell on insults, death wishes, and candidates' names, and one comment stated the function openly: avoid writing Lula's name, because the algorithm is throttling us. This is a form that reduces the chance that the platform's automatic moderation, a control coming from outside the community, will reach what is written, and

the comment that called for the disguise shows that at least one person described the contingency that way. That moderation actually throttled these comments, no datum in the case shows.

Source. On the gasoline video, contestations from both camps called the video a lie, and the replies to those who disagreed called the disagreement fake news ("essas fake news não cola aqui" ["this fake news won't fly here"]). The disqualification of the source of countercontrol, which Section 6.4 described in the 2018 groups with regard to the mainstream media, appears here on the scale of the comment, directed at whoever brings the contrary datum.

People. In the comment box there is no administrator to remove those who criticize, as in the groups observed by Evangelista and Bruno (2019), but there is the equivalent the audience can arrange: replying en masse, with mockery and insult, to the comment that disagrees. The numbers in the previous section are that consequence, counted.

Place. There was no move to a closed space; there was the opposite movement. Lula campaign comments carried hashtags of the opposing camp ("#bolsonaro2022," "#flávioBolsonaroPresidente"), which puts the message in the searches of those who follow these tags. This is a reading of function, and the case does not measure whether anyone from the other camp reached the video through them.

Was it the profile's bubble?

The opening question calls for three answers, because "bubble" covers three different things. If the bubble is the follower list, the videos did not stay in it: 99.7% of the accounts reached did not follow the profile, and the audience reached was younger and more male than the base. If the bubble is whoever watched, the case cannot say: the dashboard does not report anyone's political position, and the 2,076 comments read on the January 8 video were written by 0.2% of the accounts it reached. If the bubble is the verbal subcommunity of Section 6.1, the listeners who react to what is said and arrange consequences for it, the answer is yes, and it comes with numbers: on the January 8 video, 1.5% of comments declared the camp opposite to the video's, likes went to agreement, and the few who disagreed received replies at a rate a hundred times higher than the others, a quarter to a third of them hostile. Guerin (1992) writes that changes in language use may be under the control of verbal subcommunities smaller than the wider verbal community; this was one, assembled in two days, which did not exist as a group before the video and did not become a set of followers after it.

How an audience of a million accounts from outside the base can be so homogeneous in what it says is the question the case leaves, and it has four rival answers. The first is delivery: the platform's ranking decides how many eyes and which ones (Section 8.3), and if it uses the reactions of those who have already watched to choose who will watch next, the audience is progressively selected by its own reactions; no source in the book shows what the ranking uses as input, and neither does the case. The second is the negative audience: those from the other camp watched and did not comment, because the comment box punished those who disagreed. Skinner (1957) states that the effect of a negative audience is detected only in combination with a positive audience, as a weakening of the behavior suited to the latter, and that as long as the audiences remain separate the dilemma does not arise; the reply threads show the consequence awaiting those who disagreed, but not how many refrained from commenting because of it. The third is filtering: the 453 and 86 comments the API did not return, the filters for offensive comments, and the hidden message-request folders in Direct may have held back precisely the attacks, and this was not checked. The fourth is content: camera footage and verbatim statements, in the January 8 video, leave less to contest than a series of prices, and the gasoline video, with data that each viewer checks against his or her own memory, had nine times more comments from the opposing camp. All four may hold together, and the case does not separate them.

The case speaks to the studies of Section 6.2 at one precise point. Nyhan et al. (2023) show that content from like-minded sources makes up the majority of what Facebook users see, and that those whose exposure to it was reduced engaged more, per exposure, with the congenial content they still saw. Here, exposure did not come from the network of who follows whom, and even so the interaction was homophilic, in Cinelli et al.'s (2021) sense: those who reacted had, by an overwhelming majority, the content's leaning. That the homophily of interactions depends less on exposure than on the consequences the audience arranges is the chapter's hypothesis, and the case is compatible with it; selection by delivery, the first rival above, is equally compatible.

The five questions in Box 6.3 show how far the case goes. The correction existed and is identifiable: comments from the opposing camp and contestations with no declared camp corrected or contested the video, and the mass reply to them is the consequence the audience arranged for those who corrected. Subsequent speech was not recorded: the case is a snapshot of one night, without the before and after of those who were replied to, and it does not say whether these people commented again. The audience is not the same as at another moment, because it was assembled by the platform for these videos. Differential reinforcement was measured, in likes and replies. And the rival of approv-

al without avoidance still stands: likes and replies suffice to describe what was seen, without assuming that anyone avoided correction. Calling what happened in the comment box avoidance of countercontrol, as the chapter defines the term, would require the second question, which the case does not answer.

DATA	Two videos from a profile with 11,550 followers, published during the 2026 campaign, totaled 2.02 million views; in the month, 99.7% of the accounts the profile reached did not follow it, and the audience reached was younger and more male than the base. In the 2,952 comments read, the camp opposite to the video's was declared in 1.5% (January 8) and 13.9% (gasoline). Agreement received 91% of the likes given to comments on the first video; on the second, contestation from Lula's camp received 565 likes and contestation from Bolsonaro's camp, of the same size, 49. Comments from the opposing camp received 170 to 212 replies per hundred, and 43 to 65 hostile ones; those from the video's camp, 2 to 32 and 0 to 2. Four comments were hostile toward the author. The API did not return 18% and 9% of the comments counted.
INTERPRETATION	The commenters functioned as a verbal subcommunity assembled for each video, which reinforced agreement and responded to disagreement with aversive consequences. The absence of attacks on the author is compatible with four readings, selective delivery, the negative audience, filtering by the platform, and the form of the content, which the case does not separate.
HYPOTHESIS	If the mass reply to disagreement functions as punishment, those who were replied to will comment less afterward, and videos on topics in which each side checks the data against its own experience (prices) will attract more comments from the other camp than videos with images of the events. Testing it requires the comment history of the same accounts before and after, and comparison with profiles of other camps under the same coding rules.

For those who wish to repeat the analysis on a profile, of any camp, Box 6.2 brings together the steps the case showed to be necessary.

BOX 6.2

How to measure a profile's bubble by its consequences

1. Define the period and a baseline: the profile's usual posts, with reach and nonfollower share, before the episode.
2. Read, per post and for the account, the reach among followers and nonfollowers and the audience by age and gender; note the time of reading, because the numbers change.
3. Collect all comments and replies through the API and record how many the platform counts and how many it returns; before concluding that there were no attacks, check in the app the hidden comments, the hidden-words filter, and the hidden message requests.
4. Code position, declared camp, and hostility with written rules (Box 6.1), and separately code a random sample with a second coder, to measure agreement.
5. Measure the consequences: likes by comment category, replies and hostile replies by the commenter's camp, and the threads in which disagreement was replied to.
6. Before calling a repertoire avoidance, go through the five questions in Box 6.3; without the subsequent speech of those who were replied to, what you have is a description of consequences, not of avoidance.

6.7 Semantic analyses on social media

One of the author's books (Moreira et al., 2022a), which also brings together studies with comics and with authoritarianism (Chapter 4, Section 4.7), includes a study of semantic analysis of spontaneous posts, Study 4. The question was whether, from social media texts, it would be possible to describe equivalence classes related to the *Black Lives Matter* movement and, through them, the social view of the movement, measured by the semantic valence of the words used. Collection was automatic, through R scripts that queried Twitter's public API for 20 days, from April 22 to May 11, 2021, filtered by the movement's hashtags. Before the analysis, usernames (of those who posted and of those mentioned) were removed, as were all words that were not proper nouns, adjectives, nouns, or verbs. Those that remained were counted and classified by part of speech and by semantic valence: positive, negative, or "not applicable"; the chapter does not describe the criterion or who assigned the valence. The results are given in frequency charts, with no numbers in the text: nouns predominated, and adjectives were few; most words fell under "not applicable," and, between positive and negative words, negative ones were more numerous, including among adjectives; the most frequent words related to the George Floyd trial ("case," "jury," "decision," "justice"). Because

the collection took place in the week of the verdict in that trial, the authors conclude that it cannot be asserted that the social view of the movement is negative, even though there are more words of negative than of positive valence, and they suggest repeating the collection at the time of Floyd's death. They record difficulties, such as incomplete posts and manual separation of words.

From there, the authors propose a conceptual expansion: "shared (or common)" equivalence classes, found in many individuals, which move from an individual phenomenon to a high-frequency phenomenon and approach the concept of social representation. According to the authors, collecting spontaneous posts, without those who post knowing they would be analyzed, excludes the social desirability variable, unlike the questionnaires of Studies 1 and 2 in the same book.

For this chapter, the proposal matters for one reason and calls for one caveat. The reason is that, if the bubble is a verbal subcommunity, the natural hypothesis is that the shared classes within it are more homogeneous than those of the wider community, and the study's method, which measures the semantic valence of words in a corpus of spontaneous posts, is a way of testing it: it would suffice to split the same corpus by community (equivalence classes are discussed in Chapter 4, Section 4.2). The caveat has two parts. The study did not separate communities, because usernames were removed and collection was by hashtag, and it therefore says nothing about bubbles. And the spontaneity of posting does not eliminate the audience: whoever posts on a feed is speaking to followers, and the audience controls what is posted, as in the cycle of Chapter 9. What the study excludes, social desirability in front of a researcher, is one audience, not the absence of an audience.

6.8 Rival interpretations and limits

The chapter's reading, that the bubble is a verbal subcommunity and that part of what is maintained within it may be avoidance of countercontrol, has rivals. The box below brings them together, one by one, and the chapter's limits come after it.

Rival interpretations for the same body of data

Selection, not control. The homophily that Cinelli et al. (2021) measure may reflect who chooses whom to talk to, not what the community does with what is said: people who already agree draw closer, and the bubble would be a consequence of agreement, not its cause. The studies in the chapter do not separate the two directions.

Approval without avoidance. The repertoires of Section 6.4 may be maintained solely by the audience's positive reinforcement (likes, forwards, belonging; Chapter 9), without correction being aversive or the repertoire avoiding it. Approval suffices to explain them. Distinguishing the two contingencies requires showing that the repertoire changes after correction, and no source in the chapter records this.

Distrust as discrimination, not as avoidance. Someone who disqualifies a corrector may be under the control of a history with that source. Freeze et al. (2021) showed, in an online experiment, that invalid misinformation warnings reduce the credibility of authentic news and impair political memory. The participants were 434 people from a crowdsourcing platform (MTurk) in the United States, who watched a video of speeches by members of Congress on C-SPAN, answered unrelated questions, and read a fabricated news article about the video, which was either vague, or contained correct facts, or had some of the facts altered. Some of them, chosen at random, then received the warning that "some of the information presented in the news article you read was inaccurate," an invalid warning when the article was accurate; afterward, everyone took a recognition test of 20 facts from the video and rated the credibility of the video and the article. The warning made the article seem 0.59 points less credible on a 5-point scale, in all conditions, including when it was accurate, and the effect on memory was significant in one of three tests. If corrections err or have costs, distrusting them is not, in itself, a protective repertoire, and the analysis would need the accuracy rate of the disqualified source, which the chapter does not have.

Motivated reasoning. Bail et al. (2018) propose that exposure to messages that conflict with one's attitudes provokes counterargument and motivated reasoning. This is the authors' explanation, in cognitive terms, and they say they did not identify the mechanism. The operant reading, that what the participant declares is controlled by audience and history, is a hypothesis the experiment does not test.

Migration through use, not through moderation. Lupa reads the change of platform as migration to semi-private environments and, at the same time, warns that its results portray the universe it monitors (Agência Lupa, 2026). Evangelista and Bruno (2019) note that *zero-rating* favors WhatsApp's popularity. WhatsApp's greater presence may reflect growth in the use of the platform and of the agency's intake channel, not a flight from moderation.

One-sided samples. The Brazilian material covers, for the most part, groups of one camp (Bonow Soares et al., 2021). What is observed may be a property of any cohesive group and not of this one, and the chapter has no way of deciding.

Four limits apply to the whole chapter. First, no source records the correction and what the speaker did after it, so avoidance of countercontrol remains a hypothesis in every case. Second, exposure, interaction, and engagement are not consequences arranged by the audience, and the data on echo chambers

measure the former. Third, the Brazilian material concerns one main platform, WhatsApp, between 2018 and 2022, and comes from observation of a few groups, from monitored public groups, and from a universe selected by one agency; generalizing from it is extrapolation. Fourth, the term "avoidance of countercontrol," the grouping into four repertoires, and the two figures are proposed here, not taken from a source.

For those who want to use the chapter in the analysis of utterances (Chapters 3, 5, and 7), the consequence is practical. Asserting that an utterance is protected from correction requires a record of the correction and one of the subsequent utterance, and a fact-check, by itself, does not supply that second term. The box below brings together what to ask before calling a repertoire avoidance.

BOX 6.3

Five questions before calling a repertoire avoidance of countercontrol

1. What was the *correction*: who made it, when, with what label, and did it reach the speaker or the community?
2. Did the *subsequent* utterance change in form, source, place, or authorship relative to the earlier one, and how quickly?
3. Is the *audience* the same, with the same composition, before and after?
4. Did the audience *reinforce* the change (reactions, forwards, staying in the group) more than the earlier form?
5. Which *rival* still stands: approval without avoidance, discrimination of the source, a change of audience or of platform?

Summary

The bubble, seen as a verbal subcommunity, is a set of listeners with a shared history of reinforcement. The data on echo chambers measure something else: what reaches the feed and whom people interact with. In three studies, homophily is high on specific topics, the feed as a whole is less homogeneous than current fears assume, and neither reducing congenial exposure nor increasing exposure to the opposing camp shifted attitudes in the expected direction (Section 6.2). None of them measures who reinforces what the speaker says or how much of the correction reaches him or her, so the question of the bubble as a verbal community remains open.

Guerin (1992) gives the condition under which counterfactual talk is maintained: a small, cohesive group in which belonging is worth more than accurate reporting. Outside it, conformity pressure produces countercontrol (6.4).

The chapter's proposal is that there are repertoires of the speaker and of the community that reduce the chance or the effectiveness of this countercontrol, avoidance of countercontrol, a term proposed here, of four types: form, source, place, and people. Three weaken conditions that Delprato (2002) associates with effective countercontrol; the fourth is an addition proposed here (6.4). The Brazilian material describes each of the four, with limits of sample, platform, and political camp that Section 6.5 enumerates. What it does not provide is the correction and what the speaker did afterward; that is why avoidance remains a hypothesis, and the audience's approval, without avoidance, is the strongest rival (6.8).

Chapters 3 and 5 take from this chapter the idea that the form of speech (attitude, autoclitic, disguised mand) may be maintained because it reduces correction, and that this is tested with dated utterances, before and after fact-checks; Chapter 8, the idea that what the platform delivers to the feed decides how permeable the boundary in Figure 6.1 is; Chapter 12, the question of which interventions act on the community's contingencies, and not only on exposure.

References

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bonow Soares, F., Recuero, R., Volcan, T., Fagundes, G., & Sodré, G. (2021). Research note: Bolsonaro's firehose: How Covid-19 disinformation on WhatsApp was used to fight a government political crisis in Brazil. *Harvard Kennedy School Misinformation Review*, *2*(1). https://doi.org/10.37016/mr-2020-54

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Delprato, D. J. (2002). Countercontrol in behavior analysis. *The Behavior Analyst*, *25*(2), 191–200. https://doi.org/10.1007/bf03392057

Evangelista, R., & Bruno, F. (2019). WhatsApp and political instability in Brazil: Targeted messages and political radicalisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1434

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Guerin, B. (1992). Behavior analysis and the social construction of knowledge. *American Psychologist*, *47*(11), 1423–1432. https://doi.org/10.1037/0003-066x.47.11.1423

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Moreira, M. B., Premuli, I. T., Rabelo, E. de P., Pessoa, M. B. F., & Medeiros, M. R. G. C. (2022a). *Atitudes: Histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos* [Attitudes: Comics, semantic analyses on social media, and the stimulus equivalence paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_q0.pdf

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., . . . Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137–144. https://doi.org/10.1038/s41586-023-06297-w

Recuero, R., Tavares, C., Volcan, T., Pozzebon, M., & Dutra, M. (2025). Desinformação narrada pela checagem: estudo das eleições de 2018 e 2022 [Disinformation as narrated by fact-checking: A study of the 2018 and 2022 elections]. *E-Compós*, *28*, 1–26. https://doi.org/10.30962/ecomps.3039

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-013

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

instituto WALDEN4 | ACADEMIA WALDEN4



LAB WALDEN4

Redo the classic experiments in your browser

Classic psychology studies rebuilt to run on your computer or phone. You take part, and your data appear alongside the original results.

8 experiments to try

Free access, no sign-up required

walden4.com.br/experimentos

CHAPTER

7

Voting as Verbal Behavior

A candidate may talk to voters right up to the day before the election, but is barred from going into the voting booth with them. The observation is Lloyd's (1994), who makes it to illustrate another problem, namely that the change agent may have access only to verbal behavior and never to the behavior of interest. It serves as an opening here because it names the asymmetry that organizes the chapter: what the voter says has an audience, and the vote, because it is secret (Visser, 1996), has none at the moment it occurs.

Chapters 3 and 5 analyzed what the politician says. This chapter turns to those who listen and vote. The question is why voting can be treated as verbal behavior, that is, as behavior maintained by consequences mediated by other people, and what this changes in how we read the relation between what voters say and what they do at the ballot box. The tools are those of Chapter 3, the verbal operant and the audience, and those of Chapter 4, attitude as evaluative verbal behavior and the transfer of function among stimuli of the same class. One is missing, and it comes in at Section 7.4, at the point where the analysis needs it: the correspondence between saying and doing.

A word of warning about the sources. None of the studies used here measures the votes of Brazilian voters. Visser (1996) analyzes a Dutch election survey; Bond et al. (2012), a U.S. congressional election; the studies of attitude and behavior are, for the most part, studies of American students, from the 1930s to the 1970s. The available Brazilian studies, those of Chapter 4, deal with attitudes toward presidents, not with votes. The chapter is, therefore, a proposed analysis, and the extension of the findings to Brazil is a hypothesis, not a report of what is known about the Brazilian vote.

7.1 Voting under the definition of verbal behavior

For Skinner (1957), verbal behavior is behavior whose reinforcement is mediated by other people (the listener), while the combined behavior of speaker and listener is the "total verbal episode," in which nothing "emerges" in the social unit beyond the combined behavior of two or more individuals. The definition sets no requirement of form, mode, or medium: any movement able to act on another organism can be verbal. Pointing is verbal, because it works only if it changes what someone does.

The application to voting is made here, and it is worth saying where it is weak. Casting a vote is a movement that has an effect on a medium, the ballot or the screen, and through it on the outcome of the election, on who governs, and on what the voter's community does next. By Skinner's criterion, it is verbal to the extent that the consequence maintaining it is mediated by people. The weakness lies in the fact that the listener of this behavior is not a person present who reinforces the act, as the listener of a request is; it is a system, and reinforcement of the act, in the sense in which we defined it for verbal behavior, does not occur at the moment it is emitted. The chapter does not settle the classification by decree. It adopts a more modest claim: voting is behavior whose maintenance is social, and so the analysis of mediated controlling relations applies to it, because what maintains it is what the voter's community does with it and with what the voter says about it.

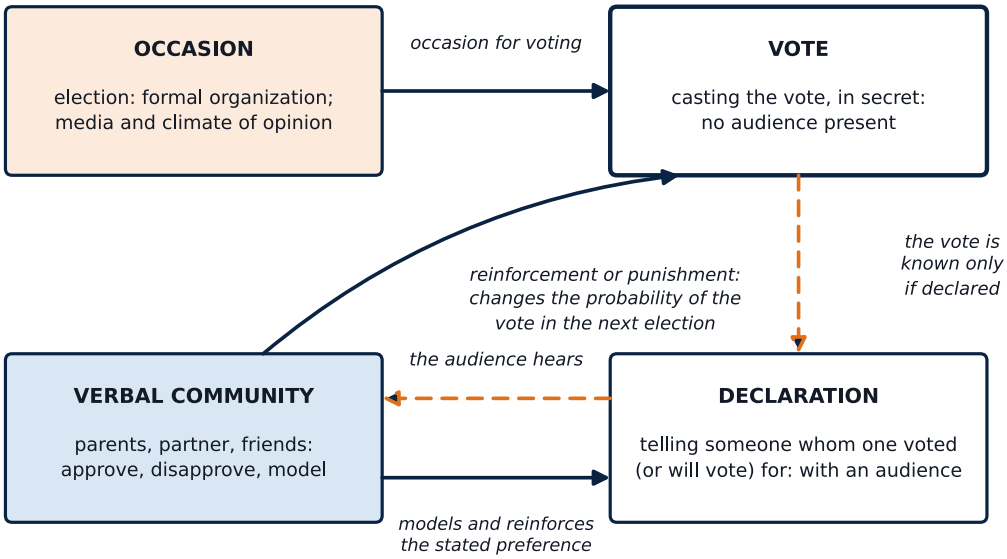
Visser (1996), whose article underpins this section, arrives at the same modest claim by another route. The author states that voting has rarely been studied from a behavior-analytic perspective, owing to strong mentalistic influences in electoral research. He starts from the premise that voting for a party is an operant response, emitted under specified stimulus conditions: the formal organization of the election is the occasion; the consequences are applied by the voter's verbal community (family, friends, colleagues, neighbors); and distal sources, such as the climate of public opinion, which the media make relevant to each voter, also specify probabilities of reinforcement or disapproval.

To test the model, Visser simplifies it. He sets aside the discriminative stimuli of the electoral situation, assuming that it is enough for the voter to know when and where the election takes place and which parties are running, and keeps two agents of consequences: a close social environment (such as the family), which models the behavior and reinforces the verbalization of a preference matching its own, and, in adult life, peers (a wider verbal community:

friends, coworkers, partners), who reinforce a vote similar to their own and disapprove of a divergent one. Figure 7.1 brings the model together and adds a box it does not include, the declaration.

FIGURE 7.1

Voting as a mediated verbal episode: occasion, vote without an audience, declaration with an audience, and verbal community



Solid: relation in Visser's (1996) model. Dashed: reading proposed here (the declaration is not in the model).

Note. Original diagram by the author, based on Visser (1996) and Skinner (1957). It does not represent data. Solid line: relation in Visser's model; dashed: the reading proposed here (the declaration is not part of the model).

The declaration box is the addition that makes the analysis possible, and Visser does not make it. In the model, social consequences bear on the vote. But the vote is secret, and what the voter's community can reinforce or punish is what it sees and hears: what voters say about their vote, before or after, and what can be inferred from what they say about other matters. Between the act and the social consequence there is, in most cases, a statement. It is through this statement that the vote enters the verbal episode, and it is on this statement that the rest of the chapter works.

7.2 Mand, tact, or intraverbal? Possible functions of the vote

As in Chapter 3, what defines the operant is what controls it, not its form. Skinner (1957) states that the mand is a type of operant defined by its controlling variables, not a formal unit: no response can be called a mand from its form alone, and identifying any verbal operant requires knowing the variables of which the response is a function. The same holds for the tact and for the intraverbal. The same vote, for the same candidate, may be under different controls in different voters, and in the same voter in different elections.

The first candidate is the tact function. The tact is the operant evoked by an object, event, or property, because, in their presence, the verbal community characteristically reinforces a response of that form (Skinner, 1957). A vote under this function would be controlled by what the candidate or the party has done, said, or represents for the voter, and would vary with the event: if what controls it changed, the vote would change. This is the most direct reading of the vote. Its weak point is that it requires contact with the event. Guerin (1994) observes that many attitudes concern events the person has never experienced, such as nuclear energy, and would therefore be intraverbal: what one knows comes from newspapers and from other people. It is reasonable to suppose that, for many voters, candidates for national office are of this kind, but there are no data on this among the sources read.

The second is the intraverbal function, the response with no point-to-point correspondence with the verbal stimulus that evokes it, of which, Skinner (1957, pp. 71–72) says, most of the "facts" of history are made. A vote under this function would be controlled by words: the party acronym, the slogan, what the family says, what the group recommends. Visser's model is, on this point, almost a description of that history: the audience reinforces the verbalization of a preference similar to its own, so that, when the election comes, the voter has a verbalized preference that the act merely continues. In this case, the vote changes when the words that evoke it and the people who say them change, not when the event changes.

The third is the mand function, or rather something that approaches it. The mand is the response reinforced by a characteristic consequence and, therefore, under the control of deprivation or aversive stimulation (Skinner, 1957). Voting to obtain or to avoid something from those who govern is the most natural reading, and the weakest: the listener is not present, has not been trained to reinforce the voter, and the reinforcer, if it comes, comes to the whole community and after a long delay. The mand function fits the declaration better than the vote. Guerin (1994) shows the mechanism for attitude statements: a

boss who says she does not like people who drink coffee at work is not making an innocent tact of her private attitude, and functions as if she had manded "Do not drink coffee while you work or else!", whether or not she intends it. Someone who says "I'm voting for so-and-so" in front of someone who has not yet decided may be doing just that, and the rule of Chapter 5 applies here: form does not decide.

There is a fourth possibility, which is not one more verbal operant but changes what can be expected of the vote: the rule. Street (1994), citing a 1980 text by Craig, gives the example of the voter who says he is a Democrat and votes for a Republican for president. Traditional research takes this as attitude inconsistency; for Craig, it may be only an artifact of the researcher's failure to identify the salient aspects of the stimulus in light of the person's history, because the voter may follow a rule of never voting for a presidential candidate of the party in control of Congress. The "inconsistency" then measures the observer's failure to devise effective verbal probes in light of the speaker's history. A vote that follows from a rule is rule-governed behavior, in the terms of Section 3.7, and may be neither tact, nor intraverbal, nor mand.

What can be said, given the state of the sources, is little. The four possibilities are candidates, and none has been tested for the vote of a Brazilian voter. The only operant model of voting among the sources is Visser's, which treats the vote as a response maintained by social consequences and sets aside the discriminative stimuli of the election. What the analysis gains from the list is a precise question for each case: what would change this voter's vote, and who would know it.

7.3 The secret ballot as removal of the audience

The audience is, for Skinner (1957), the discriminative stimulus in whose presence verbal behavior is characteristically reinforced. Audiences differ in what they do with verbal behavior: one that merely does not reinforce is distinct from one that punishes, and an audience in whose presence verbal behavior is punished is a negative audience. The confidant is the example of the non-punishing audience, the sympathetic person to whom one can talk while fearing punishment less than with listeners at large, and, in that person's presence, the speaker may emit behavior that other audiences punish. Chapter 10 (Section 10.1) deals with these audiences in full; here two ideas are enough: the audience is part of the occasion for verbal behavior, and what one audience punishes may be emitted in the presence of another.

The secrecy of the voting booth is, in these terms, the institutional removal of any identified audience at the moment of the act. Voters cast their votes without anyone who approves or disapproves of them being present, and without the act, at that moment, producing any social consequence. The chapter's hypothesis, which no verified source tests, is that this shifts control: the audience ceases to be the occasion for the act, and what remains is what the voter's history has left behind, including rules and verbalized preferences learned in the presence of those audiences. What the removal does not eliminate is what voters said before and what they will say afterward. The declaration of a vote, unlike the vote itself, is emitted in the presence of an audience, and it is under that audience's control that one expects to find it.

There is a consequence for researchers, which Visser (1996) notes candidly. Because the vote is secret, the researcher cannot inspect it and relies on voters' own verbal reports about their voting behavior and about the factors that reinforce it. These reports may be controlled by contingencies other than those of the vote they are supposed to reflect, especially those of the interview situation in which survey data are typically gathered. Visser then assumes that, in a survey interview that tries to be unobtrusive, the report accurately reflects voting behavior and the factors that control it. It is a working assumption, declared as such, and the rest of the article depends on it.

Guerin (1994) provides the counterpoint. Survey scales and questions depend on who is asking and on the verbal culture in which the survey is conducted, because attitudes always depend on social consequences; and he cites, at this point, a 1986 study on the presence of others and the reporting of voting in American elections. That study was not read; only the reference is known, and no claim is made about what it found. For the chapter, what Guerin draws from it is enough: the report of a vote, like any other report, depends on who is asking.

DATA	Because the vote is secret, the researcher does not inspect it and uses verbal reports, which may be under the control of different contingencies than those of the vote, in particular those of the interview (Visser, 1996). Survey scales and questions depend on who is asking and on the verbal culture in which the survey is conducted (Guerin, 1994).
INTERPRETATION	What electoral research observes is the voter's declaration under the audience of the interview, not the vote. The relation between the two is the object of study, and it cannot be presupposed.

HYPOTHESIS That the public declaration of a vote is more under the control of the audience than the vote is, is a reasonable and untested hypothesis: there is no study among the verified sources that compares, for the same voters, what they declare before different audiences with their validated vote.

7.4 Declaring and voting: verbal attitudes and acts

The question of the relation between what people say and what they do is older than electoral research, and behavior analysis is not the only field to have asked it. Lloyd (1994) starts from the common-sense acknowledgment that the correlation between what people say they would do and what they actually do later is low, or even nil, and notes that the psychologist, who often relies on verbal report instead of observing the criterion behavior, is less skeptical than the layperson. This section brings together what the verified sources show, and begins with the concept that organizes it.

CONCEPT IN USE

Say-do correspondence

Say–do correspondence is the relation between what people say they did, or will do, and what they actually do. In Risley and Hart (1968), twelve children aged 4 to 5, from low-income families, in a half-day experimental preschool, formed two groups of six. The teachers recorded the use of materials during the morning free-play period (the doing) and, about an hour and a half later, at the last snack before the children went home, asked each child what good things he or she had done that day (the saying). First, the snack and approval were contingent on reporting the use of a specific material: blocks in Group A, paint in Group B. With the snack depending only on the report, the proportion of children reporting rose to a median of 100% in both groups, with no reliable effect on use. Then the snack was made contingent on the report corresponding to the observed use, and correspondence rose, by different routes: in Group A, use increased to the level of the report (a median of 100% for both saying and doing); in Group B, reporting declined (to a median of 40%) to the level of those who had actually painted (25%).

Two consequences matter here. The first is that saying and doing are, in Lloyd's (1994) words, separate response classes that may or may not develop independently, and nothing implies that one causes or influences the other. The second is that correspondence is not a property of the person: it is the effect of a contingency that reinforces it. Reinforcing saying alone does not produce doing.

Risley and Hart started from earlier studies that had reinforced only verbal behavior and found little increase in related nonverbal behavior. The same study opens, in Lloyd (1994), the section on do–say and say–do, and it is through it that the concept enters the discussion of attitudes. Table 7.1 brings together five verified sources: two, which this text knows through Lloyd's account, show the gap in a field study and in a review of 31 studies; two measure it in a concrete request made to the participant; and one shows a condition that reduces it.

TABLE 7.1

Five sources on the gap between what people say and what they do, and one condition that reduces it

Source	"Saying"	"Doing"	Result	What it does not allow
LaPiere, 1934, as reported by Lloyd (1994)	Proprietors' written answer about serving Chinese customers	Service to a Chinese couple, on visits to about 250 establishments	About 90% of the 128 who answered said they would not serve them	Checking the original study, which was not read
Wicker, 1969, a review reported by Lloyd (1994)	Attitude scales	Measures of overt behavior, in 31 studies	Correlations rarely above +0.30, many near zero, negative in four	Checking the 31 studies; speaking of voting
DeFleur and Westie (1958)	Summated differences scale; 23 more and 23 less prejudiced students, matched	Signing, on a form with graded uses, permission to be photographed with a Black person of the opposite sex	The prejudiced avoided more, but 14 of 46 did the opposite of what was predicted; 71.8% invoked a reference group	Generalizing: 46 students; the authors did not test the mechanism
Linn (1965)	Verbal attitude (prejudice)	Posing for a photograph with a Black person of the opposite sex	Discrepancy in 59% of cases; relation a function of social involvement and prior experience	Estimating the rate in other populations
Regan and Fazio (1977)	Attitude, formed through direct experience or through information	Related behavior, in the laboratory and in the field (housing crisis)	Greater consistency with direct experience	Speaking of public declarations; of voting

Note. Figures from the articles and, in the first two rows, from Lloyd's (1994) account. The studies by LaPiere and Wicker are known here only through that account and, in LaPiere's case, through a mention in DeFleur and Westie (1958); the pages cited are Lloyd's. "Measures of behavior" in Wicker are those the 31 studies used, not voting.

The first two rows show how old the question is. LaPiere traveled across the United States with a Chinese couple, who were served in over 250 restaurants, hotels, and similar establishments, with a single refusal (DeFleur and Westie,

1958); when he later asked, in writing, whether they would accept Chinese customers, about 90% of the 128 proprietors who answered said they would not (Lloyd, 1994). Wicker reviewed 31 studies that measured attitude with scales and overt behavior: the correlations rarely exceeded +0.30, many were near zero, and in four studies they were negative (Lloyd, 1994). A correlation of +0.30 is weak: squared, it indicates that less than a tenth of the variation in one measure goes along with the variation in the other. The number that matters for the chapter is Wicker's: the gap between saying and doing is not the exception. The seriousness of the number, Lloyd writes, is greater when one notes the widespread acceptance, among both laypeople and psychologists, of the notion that, to change behavior, one must first change attitudes, a notion that requires a high positive correlation.

The next two rows are studies in which the doing was a decision in response to a concrete request, and what is of interest is what the authors say about the source of the discrepancy. DeFleur and Westie (1958) administered prejudice scales to 250 students and selected 23 from the most prejudiced quartile and 23 from the least prejudiced, matched on eight characteristics. After a laboratory session and an interview, they asked each one to sign, on a release form with a graded series of uses, permission to be photographed with a Black person of the opposite sex. The prejudiced avoided this more, but 14 of the 46 showed behavior opposed to their verbal attitude, and the relation, although significant, was not a one-to-one correspondence. In 71.8% of cases the subjects invoked some reference group when deciding to sign or refuse, and all of those who did so mentioned some type of peer group; the authors attributed the lack of a linear relation to social constraints and suggested that attitude scales may need a systematic categorization of them.

Linn (1965) repeated the design with an attitude measure identical to the act. Thirty-four female undergraduates at Wisconsin, aged 18 and 19, had answered a questionnaire with seven items of the type "I would be willing to pose with a Negro of the opposite sex if the picture were to be used" in a given context, ranging from laboratory use to a nationwide campaign leaflet for racial integration. Four weeks later, without knowing they were taking part in a study, they were invited to help construct a personality test and signed, or did not sign, the same seven releases. In 59% of cases, the verbal attitude and the number of releases signed differed by two or more points out of seven. Linn concluded that how prejudice relates to discrimination depends on the degree of social involvement with the object and on prior experience with it. In operant language, social constraints and reference groups are audience (Sections 7.1 to 7.3).

The condition that reduces the gap, in turn, is direct experience. Regan and Fazio (1977) conducted a laboratory experiment and a field study. In the laboratory, 28 Cornell freshman volunteers were randomly assigned to two conditions: in the direct experience condition, they worked through the examples of five types of problems; in the indirect experience condition, they saw the same examples already solved and explained by the experimenter. All rated their interest in each type, from "extremely boring" to "extremely interesting," which measured attitude, and then had 15 free minutes to work on whichever problems they chose; the order and the proportion attempted of each type measured behavior. Consistency between attitude and behavior was greater in subjects whose attitude had been formed through direct behavioral experience than in those whose attitude had been formed through information from the experimenter. In the field study, questionnaires were sent to freshmen whom the university had placed in temporary housing for lack of space (direct experience of the housing crisis) and to freshmen placed in permanent housing from the start (indirect experience); they measured attitudes about the crisis on seven items and offered six possible actions, such as signing a petition or writing a letter, and consistency between attitude and action was greater in the students with direct experience. For voting, the question is what direct experience with the object would be: for many voters, the candidate arrives through words (Section 7.2), and the direct experience voters have is that of voting and of declaring their vote.

For behavior analysts, the problem lies not only in measurement. Street (1994) writes that, in the behavior-analytic view, attitudes are the behaviors observed when someone is asked to complete an attitude survey, with no lasting mental state being captured, and that responding to a scale is behavior emitted under unusual social conditions. The cognitivist assumes that self-report and behavior in the presence of the object are effects of the same cause, the attitude; the behavior analyst, that each is produced by its own causal events and that, if they are different, the behaviors will probably be different. The criterion of consistency is itself verbal, and self-reports are reinforced by others for being plausible, not for being objectively accurate. Guerin (1994) arrives at the same place: given the multiple sources of control, the broad repertoires, and the various verbal communities, we should expect inconsistencies between attitude and behavior, especially when the behavior is verbal.

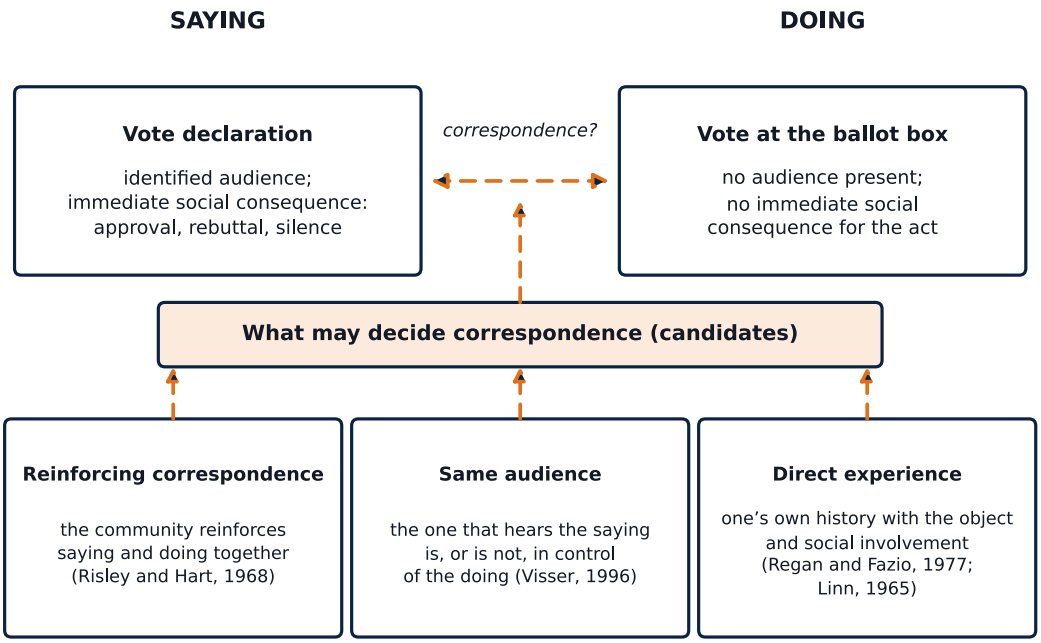
Figure 7.2 reduces the section to one question: what decides whether what voters say and what they do correspond? The three candidates come from the sources: the history of reinforcement of correspondence (Risley and Hart, 1968); the identity, or lack of it, between the audience of the saying and the one that controls the doing, a reading that follows from Visser (1996) and from

DeFleur and Westie (1958); and direct experience with the object (Regan and Fazio, 1977; Linn, 1965). None has been tested for vote declaration and vote in the same voter.

Lloyd (1994) adds two observations about order, both from studies that were not read. In a 1973 study by Israel and O'Leary, children who played for 20 minutes and, 40 minutes later, said what they had played with (do–say) showed greater correspondence than those who said what they would play with and, 40 minutes later, played (say–do). And, in a 1977 study by Sowers and colleagues, college students' reports of the studying they had done 14 hours earlier corresponded more closely to that studying than their plans corresponded to the studying done 6 hours later. The chapter's reading is that declaring after voting, as in a declaration collected at the exit of the polling place, is a do–say response, and declaring beforehand is a say–do response.

FIGURE 7.2

Saying and doing: what may decide whether the vote declaration and the vote correspond



Original diagram, no data. Dashed: conceptual relation or hypothesis.

Note. Original diagram by the author, based on Risley and Hart (1968), Visser (1996), Regan and Fazio (1977), and Linn (1965). It does not represent data. Dashed line: conceptual relation or hypothesis.

DATA	In 31 studies reviewed by Wicker in 1969, correlations between attitude scales and measures of behavior rarely exceeded +0.30 (Lloyd, 1994). In DeFleur and Westie (1958), 14 of 46 students acted in opposition to their verbal attitude; in Linn (1965), the discrepancy was 59%. In Risley and Hart (1968), reinforcing the report raised the report without changing use, and reinforcing correspondence produced it in both groups, by different routes.
INTERPRETATION	The gap between saying and doing is not an anomaly to be explained by measurement error: it is what one expects of two response classes with their own contingencies (Street, 1994; Guerin, 1994). When there is correspondence, it seems to depend on a history of reinforcement of correspondence itself (Risley and Hart, 1968) and on direct experience with the object (Regan and Fazio, 1977). Nothing in the sources concerns voting.
HYPOTHESIS	That correspondence between vote declaration and vote is greater when the same audience takes part in both moments, and smaller when the audience of the declaration differs from the one that controls the vote, is a hypothesis proposed here, and no source tests it. That declaring after voting corresponds more closely than declaring beforehand, as Lloyd reports for children and college students, is testable with validated votes and declarations collected at different times, and there is no study among the sources read that does so.

7.5 Social consequences of voting: the field experiments

If the analysis of voting as verbal behavior is correct, what the voter's community does with the vote, or with what is said about it, changes the probability of voting and of voting the same way. Two sources deal with this, with opposite designs: a survey, which can only associate, and a field experiment, which manipulated.

The survey is Visser's (1996), with data from the 1986 Dutch National Election Study, a two-wave survey, before and immediately after the election, and the only one in the series that measured the votes of the people close to the voter. Visser classifies each voter according to the reinforcement history the model assumes, that is, according to whether the parents vote for the same party and whether there is a partner whose party is known, and checks whether the voter's vote matches the partner's or, with no identifiable partner, the parents'; in that case, the direction of the vote is considered explained by the model. Among 1,255 voters, the model explains the direction of 812 votes

(65%), against 41% expected if voters were randomly distributed among the categories. As for consistency, which Visser defines as how often the voter chooses the same party in successive elections, votes modeled or reinforced by parents or partner were more consistent than those disapproved of by significant others, generally with a large and significant difference: 68% (parents and partner voting the same way) and 67% (parents only) of voters said they had always voted for the same party, against 29% to 36% in the cells in which the voter diverged from parents or partner. The exception the author highlights is the cell in which parents and partner voted differently and the voter voted like the partner: 32%, which, for him, underscores the importance of reinforcement history. Visser cautions that survey research does not have the controlled character of experimental studies, so that causal claims call for caution, and that analysis at the level of the individual voter may miss the impact of macrosocial developments; and he concludes that, from a behavior-analytic standpoint, there is no valid reason to grant voting a privileged status among human actions.

The experiment is Bond et al.'s (2012), with 61 million Facebook users in the 2010 U.S. congressional elections. Users were assigned at random to a social message (60,055,176 users: encouragement to vote, a link to the polling place, an "I Voted" button, a counter, and up to six photos of friends who had clicked the button), to an informational message (611,044: the same, without the photos), or to no message (613,096). They measured three things. Clicking "I Voted" is a measure of political self-expression, which, according to the authors, probably depends on how much users want others to see them as voters; clicking the polling-place link measures information seeking, and is not reported to the social community; and voting was validated in public records for 6.3 million users.

There are three results. Those who received the social message clicked "I Voted" more than those who received the informational one (20.04% versus 17.96%), and voted 0.39% more than those who received no message and 0.39% more than those who received the informational one; those who received the informational message turned out at exactly the same rate as the controls. Each close friend who had received the social message made users 0.224% more likely to vote (validated vote) and 0.099% more likely to express having voted; there was no evidence of an effect of other friends on validated voting or on polling-place searches. And the authors estimate that the social message directly added about 60,000 voters to turnout and, through social contagion, another 280,000, for a total of 340,000 votes, about 0.14% of the voting-age population; all effects are intent-to-treat effects, that is, calculated over everyone assigned to the message, including those who had already

voted early or logged on to Facebook too late to vote; they therefore underestimate the effect on those who could actually be treated. Almost all of the transmission occurred between close friends, who probably have a face-to-face relationship.

In operant terms, the social message differs from the informational one in a single respect: the photos of friends who had declared that they voted. The experiment therefore isolates the fact that the social datum, and not the information about where and when to vote, added turnout. What, within the social datum, was at work, the design does not separate. Three readings are compatible: that of the model (seeing friends vote); that of the audience (the anticipation that friends will see one's own "I Voted," the self-expression the authors measure); and that of the rule (seeing that people in the group vote makes applicable a rule the voter already had). Saying and doing were measured under the same manipulation and varied together, in the sense that the social message increased both clicks and validated voting. What the article does not show is correspondence for each user: validated voting was studied through a group-level procedure, described in the supplementary material, and one cannot see, for each person, whether the click corresponded to the vote.

There are other field experiments on social pressure and turnout, and they are not cited because their full text could not be read through legal means. The gap belongs to the body of sources read, not to the literature. For this reason, too, the chapter does not claim that the effect found by Bond et al. extends to elections in other countries, to other platforms, or to other years.

DATA	Among 1,255 Dutch voters, votes modeled or reinforced by parents or partner were more consistent across elections than those disapproved of, and the direction of the vote was explained in 65% of cases, against 41% expected by chance (Visser, 1996). Among 61 million American users, the message with friends' photos increased clicks on "I Voted" and validated voting (0.39% more than with no message); the message without photos did not change turnout (Bond et al., 2012).
INTERPRETATION	Both are compatible with the reading that what the voter's community does with the vote, or with what is said about it, changes the probability of voting and of voting the same way. Bond et al.'s is a manipulation, on a single platform; Visser's is an association, and the author himself warns about causality.

HYPOTHESIS That the mechanism is friends' anticipated approval or disapproval (social reinforcement) is a hypothesis, and the design does not separate it from modeling and from rules. That the same holds for Brazilian elections is an extension, not data.

7.6 Candidates' names and numbers in equivalence classes

Candidates' names, party acronyms, and numbers are the stimuli with which voters arrive at the ballot box, and Chapter 4 (Section 4.5) left a hypothesis about them: that the right–left label is part of the class with which the politician's name reaches the voter. What Chapter 4 had to support it is little. In the study by Pereira and Moreira (2021b), five psychology students learned, by trial and error, which abstract figures went with each of three photographs of Brazilian presidents (Bolsonaro, Lula, and Fernando Henrique Cardoso). The party acronyms (PSL, PT, and PSDB) and the words right, left, and center were correctly related to the abstract figures without specific training, which the authors take as a suggestion that they were already part, before the experiment, of a class with the presidents' photographs; and, after learning the relations between photographs and abstract figures, the ratings of the figures, made on a scale of opposite adjectives, came closer to those of the corresponding photograph. In another study by the same authors, with two experiments of five participants each, the increase in the similarity between ratings (100 minus the absolute difference between them) was, with neutral faces of unknown people, half of that obtained with abstract figures, which they attribute, probably, to the faces' prior affective load (Pereira and Moreira, 2021a). The authors say they know of no study of attitudes using socially meaningful stimuli linked to political positions.

There is a caveat about results, which Chapter 4 has already given. Haydu et al. (2015) taught 28 men from organized fan groups of three Brazilian soccer clubs, through matching-to-sample, to relate the clubs' emblems to abstract paintings and those paintings to the words good, bad, and regular. The tests assessed whether relations would emerge, without direct teaching, between the emblem of the club supported and the word bad, between that of the main rival and the word good, and between that of the minor rival and the word regular, so that the words conflicted with each fan group's history. The participants formed the trained (baseline) relations and the symmetry relations, but not the transitivity and equivalence relations between emblems and words, and the most frequent errors were choosing what the history prescribed: good for the emblem of the club supported and bad for that of the main rival. What the study shows, for this chapter, is that the intended class

does not always form when it runs counter to the participant's history. The finding is not generalized to party affiliation, and neither is that of Pereira and Moreira.

The chapter's hypothesis is as follows. In any election in which the choice requires recognizing or producing a candidate's name, acronym, image, or number, the act presupposes a conditional discrimination among these stimuli, and the analysis of equivalence classes deals with stimuli arbitrarily related through history, which become substitutable, with the function of one reaching the others without direct contact (Section 4.3). If the candidate's name, image, acronym, and number are part of the same class for the voter, the evaluative function of any one of them may reach the others, and the vote may be less under the control of what the candidate said (Chapters 3 and 5) than of the relations the voter has learned among the stimuli that designate the candidate. There is no source read on the voting procedure used in Brazil, and so it is not described: the hypothesis holds for any procedure that asks the voter for one of these stimuli.

What would test the hypothesis is a study that measured, for candidates from different political camps and with the same procedure for all, the relations among name, image, acronym, and number (by matching-to-sample or by a naming task), the evaluative function of each (by semantic differential, as in Chapter 4), and the transfer of that function after a history of matching. The symmetry rule applies: same procedure, same criteria, for all camps.

DATA	Party acronyms and the words right, left, and center were correctly related to abstract figures, without specific training, by five students (Pereira and Moreira, 2021b). Among 28 soccer fans, baseline and symmetry relations formed, but not transitivity and equivalence between emblems and words (Haydu et al., 2015).
INTERPRETATION	Transfer of evaluative function among photographs, acronyms, and figures is possible in the laboratory and depends on prior history with the stimuli; where what is taught runs counter to the participant's history, the intended class did not emerge (Haydu et al., 2015).
HYPOTHESIS	That the candidate's number, like the name and the acronym, enters a class with the photograph and with the right–left label, and that this mediates the vote, is a hypothesis proposed here. No verified source has manipulated candidates' numbers or measured the vote in equivalence studies with voters.

7.7 Rival interpretations and limits

The chapter's reading, that voting is behavior whose maintenance is social and whose relation to what voters say depends on contingencies specific to each, has rivals, and Table 7.2 brings them together with what would separate them from it.

The chapter's limits are those of its sources. None measures the votes of Brazilian voters: the study of voting in operant terms is a 1986 Dutch survey, for which the author himself flags the question of causality, and the experiment is from one platform, in one country, in one election. The attitude–behavior literature measures attitude, not voting, and LaPiere and Wicker are known here only through Lloyd's account. The classification of voting as verbal behavior is proposed here, and it stands as analysis, not as a finding. And no source compares political camps, so the chapter says nothing about differences between voters from different camps.

TABLE 7.2

Rival readings of the vote and of what voters say about it: what would support them and what would separate them

Rival reading	What would support it	What would separate it from the chapter's reading
Voting is nonverbal behavior, with no social mediation; the declaration is another response, and the relation between them is not a question about voting	Votes that do not vary when social consequences change, or after declarations	Bond et al.'s social message changed validated voting; Visser's model explains 65% of votes
Shared environment: parents and partner vote alike because they share information, neighborhood, and income, and consistency results from this, not from reinforcement (Visser, 1996, on causality)	Equal consistency among voters who are neither approved nor disapproved of by parents or partner	A design that holds the environment constant and varies approval, or that follows the voter over time
Attitude as disposition: declaration and vote are effects of the same cause, and inconsistencies come from poor measurement (described by Street, 1994)	High correlations between declaration and validated vote with better instruments	Saying and doing varying for different reasons (Risley and Hart, 1968; DeFleur and Westie, 1958)
Reminder and information: the social message merely reminded people to vote	The same effect with the informational message	Bond et al.: the informational message had the same turnout as the control; but the experiment does not separate model, audience, and rule
Inaccurate report: Visser's consistency comes from memory and social desirability, not from behavior	Reports of past votes that come close to the report of the present vote	Votes validated in public records, as in Bond et al.

Summary

Voting is behavior whose maintenance is social, and so the analysis of verbal behavior applies to it, even though the listener of the act is a system and not a person present (7.1). The same vote may be under the control of tact, intraverbal, mand, or rule, and form does not decide (7.2). Secrecy removes the identified audience from the act and leaves the declaration, which has one; what electoral research observes is the declaration, under the audience of the interview, not the vote (7.3). The relation between saying and doing has been, since LaPiere, low and variable, and behavior analysis reads it as that of two response classes with their own contingencies; correspondence is the effect of a contingency that reinforces it, and reinforcing saying alone does not produce doing (7.4). There are associational data and experimental data showing

that what the community does changes the probability of voting, but the design does not separate model, audience, and rule (7.5). The hypothesis that a candidate's name, image, acronym, and number form a class for the voter is proposed here, and no source tests it (7.6).

Chapter 10 (Section 10.1) deals in full with the negative and nonpunishing audiences on which Section 7.3 draws, and Chapter 12 returns to say–do correspondence when evaluating interventions whose outcome is what the participant says.

References

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

DeFleur, M. L., & Westie, F. R. (1958). Verbal attitudes and overt acts: An experiment on the salience of attitudes. *American Sociological Review*, *23*(6), 667–673. https://doi.org/10.2307/2089055

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Haydu, V. B., Camargo, J., & Bayer, H. (2015). Effects of preexperimental history on the formation of stimulus equivalence classes: A study with supporters of Brazilian soccer clubs. *Psychology & Neuroscience*, *8*(3), 385–396. https://doi.org/10.1037/h0101276

Linn, L. S. (1965). Verbal attitudes and overt behavior: A study of racial discrimination. *Social Forces*, *43*(3), 353–364. https://doi.org/10.2307/2574765

Lloyd, K. E. (1994). Do as I say, not as I do. *The Behavior Analyst*, *17*(1), 131–139. https://doi.org/10.1007/bf03392658

Pereira, P. B. G., & Moreira, M. B. (2021a). *Atitudes em relação a presidentes do Brasil: Formação de classes de equivalência e transferência de função* [Attitudes toward Brazilian presidents: Equivalence class formation and transfer of function] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/pereira_moreira_2021_presidentes.pdf

Pereira, P. B. G., & Moreira, M. B. (2021b). Transferência de atitudes em relação a presidentes da república [Transfer of attitudes toward presidents of the republic]. *Acta Comportamentalia*, *29*(4), 137–154. https://doi.org/10.32870/ac.v29i4.80318

Regan, D. T., & Fazio, R. (1977). On the consistency between attitudes and behavior: Look to the method of attitude formation. *Journal of Experimental Social Psychology*, *13*(1), 28–45. https://doi.org/10.1016/0022-1031(77)90011-7

Risley, T. R., & Hart, B. (1968). Developing correspondence between the nonverbal and verbal behavior of preschool children. *Journal of Applied Behavior Analysis*, *1*(4), 267–281. https://doi.org/10.1901/jaba.1968.1-267

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Street, W. R. (1994). Attitude-behavior congruity, mindfulness, and self-focused attention: A behavior-analytic reconstruction. *The Behavior Analyst*, *17*(1), 145–153. https://doi.org/10.1007/bf03392660

Visser, M. (1996). Voting: A behavioral analysis. *Behavior and Social Issues*, *6*(1), 23–34. https://doi.org/10.5210/bsi.v6i1.278



PART III

The Ecology of Information

8 Attention Economy: The Algorithm as a Programmer of Contingencies

9 The Producer-Audience Cycle

10 The Overton Window: What Can Be Said Without Being Punished

11 Infodemiology: How Information Spreads and What Is Measured

CHAPTER

8

Attention Economy: The Algorithm as a Programmer of Contingencies

Upworthy, a viral news site, used to test headlines in experiments, and one of them serves as a starting point, although it will be better addressed further on. The same story was shown to readers with four different headlines, and the click-through rates were 0.94%, 0.79%, 0.60%, and 0.58% (Robertson et al., 2023, Table 2). With the same story and the same outlet, the most-clicked headline received about 1.6 times the clicks of the least-clicked one (our own calculation). The site did this serially: the dataset comprises 22,743 randomized tests, with about 105,000 headline variations.

Each test is a clean experimental arrangement: the headline varies, the story and the audience are the same, and what is measured is the attention each variation obtains, counted in clicks. What happens to what is said when this is how attention is distributed? This chapter deals with the economy in which that distribution takes place. Section 8.1 presents attention as a reinforcer; Section 8.2, the content properties that the data associate with more clicks, shares, and reactions; Section 8.3, the rule by which the platform decides how many people see each emission; Section 8.4, the hypothesis that this delivery functions as intermittent reinforcement; and Section 8.5, the experiments in which the rule was manipulated.

The chapter is cautious on two points that current discussion takes as settled. The first is that the feed is a variable-ratio schedule: no verified source measures any schedule, and Section 8.4 states what the schedules literature describes and what would be needed for the comparison. The second is that the

algorithm rewards anger or favors one political side: what the data show is narrower, and Section 8.3 describes it in the authors' terms. In every section, data, interpretation, and hypothesis are kept separate.

8.1 Attention as a reinforcer

Skinner (1957) treats verbal behavior as behavior maintained by consequences mediated by listeners, and the listener's attention appears in scattered passages of the book, the most explicit on p. 199. The box brings together what he says and what this chapter takes from it.

CONCEPT IN USE

Attention as a generalized conditioned reinforcer (Skinner, 1957)

Chapter 3 (Section 3.3) presented the generalized conditioned reinforcer: the event that characteristically precedes many different reinforcers and that can therefore strengthen behavior whatever the relevant state of deprivation or aversive stimulation, such as approval (Skinner, 1957). The listener's attention has the same structure: it is the common consequence of generalized mands such as *Say!* and *Here!*, and, because what the listener does next can matter under many different deprivations, these responses come under wide-ranging control. Skinner gives it as an example of a variable that strengthens verbal behavior without respect to form: whatever verbal behavior draws the listener's attention is reinforced, independently of anything else the listener specifically does. Guerin's (1994) generalized social reinforcement, which Chapter 3 also presented, is a neighboring concept: attitudes are statements that a verbal community maintains through generalized social reinforcement.

Two points in the box matter for the rest of the chapter. The first is the reason attention is a reinforcer of such broad scope: it is a condition for the listener to do anything else for the speaker, so it serves many states of deprivation at once. The second is the clause "without respect to form." What is strengthened is any response that evokes attention, and attention, as a consequence, does not by itself distinguish the true from the false, nor what informs from what irritates. This last sentence is a reading proposed here on the basis of Skinner, who does not deal with disinformation or platforms. It is this reading, however, that makes it meaningful to ask, in the following sections, which content properties attract attention and who controls its distribution.

On a platform, the listener's attention is not observable to the speaker. The producer sees counts: number of followers, likes, shares, comments, clicks. The finest-grained measure among the verified sources is Twitter's "linger impression," the record that at least 50% of the area of a post was visible for at least 500 milliseconds, which Huszár et al. (2022) call the best available proxy

for whether the user was exposed to the content. It is a measure of exposure, not of attention in the functional sense; the rest of what the sources report are explicit reactions. This chapter calls the count of reactions that the platform returns measured attention. The gap between it and the attention that reinforces, the listener's, is a premise of the reading, not a finding.

One study directly asks how much attention is worth among the reasons for using social media. Gunnarsson et al. (2026) gave 328 students at a Canadian university (mean age 22.7 years; 68% women) a hypothetical social media purchase task. Participants imagined that they could use social media only by buying time, sold by the minute, and said how many minutes they would buy per day at each of 15 prices, from $0.01 to $1,120 per minute, in four function conditions: using social media solely to escape boredom and unpleasant feelings (escape), to seek information (tangible), to get engagement with their own content, such as likes, shares, and comments (attention), or to feel good (automatic). Function affected consumption at minimal cost (Q_0, what would be bought if time cost nothing), but not price sensitivity (α, how quickly purchases fall as price rises), and, as the authors' abstract puts it, escape, information seeking, or automatic positive reinforcement produced higher demand than the pursuit of likes. In the attention condition, 57.6% of individual trajectories failed the ΔQ systematicity criterion, one of the quality indicators for demand data (13.4%, 20.7%, and 11.6% in the other conditions), mostly because of zero consumption, and 51.5% of participants reported only zero values (7.0% in the tangible condition). The authors read this as low demand for using social media to obtain attention, not as inattention to the task, but they caution that the task is hypothetical.

What the study tells the chapter is a caution. If attention were the dominant reinforcer for social media users, one would expect high demand precisely in the condition named after it, and that is not what appeared, in a student sample and in a task that isolated one function, although many people use social media for multiple, overlapping functions (a limitation the authors acknowledge). There is also a difference in construct: the study measures the value, to the user, of using social media to obtain attention, not the effectiveness of attention as a consequence of a specific post, which is the chapter's question. The two can diverge.

DATA	Skinner (1957) describes the listener's attention as the common consequence of generalized mands and as a variable that strengthens verbal behavior without respect to form; in a hypothetical task with Canadian university students, demand for using social media to obtain attention was lower than demand for escape, information, or automatic positive reinforcement (Gunnarsson et al., 2026).
INTERPRETATION	The listener's attention is a generalized reinforcer of broad scope for verbal behavior. On a platform, what reaches the producer is a count of reactions, which serves as an approximate measure of that attention.
HYPOTHESIS	The count of reactions is the reinforcer that maintains what the producer posts and is dominant over other reinforcers. The studies in Chapter 9 give partial support to the first half (likes and posting frequency; social feedback and outrage), with small effects and specific populations; the second half has not been shown, and the result of Gunnarsson et al. (2026) calls for caution.

8.2 What the data show about emotional, moralized, and negative content

Three verified studies link content properties to audience reaction. They measure different responses (clicking, sharing, posting), and the difference is part of what there is to say.

Robertson et al. (2023) analyzed the Upworthy tests, in which each visitor to the site was randomly shown one of several headline variations of the same story, and the click to read the story was recorded (visitors between 2013 and 2015). The analyzed set comprises 22,743 tests, of which 12,448 passed the preregistered filter, which discards, for example, tests with a single headline (many tested teaser images rather than headlines). These tests compared an average of 4.31 variations each, totaling 53,699 headlines, more than 205 million impressions, and 2,778,124 clicks; the average rate was 1.39%. The words in the headlines were classified one by one by a language dictionary, LIWC, with 620 positive and 744 negative words. In a headline of average length, every extra negative word raised the click-through rate by 2.3%, and the presence of positive words decreased it by about 1.0%. The authors point out that positive words were slightly more prevalent than negative ones, 2.83% versus 2.62% of words. The effect of negative words was strongest for Government and Economy stories. Among discrete emotions, sadness increased the likeli-

hood of a click and joy decreased it; anger had no statistically significant effect, and fear had an effect in the direction opposite to the preregistered one.

The design has limits that the authors themselves state. Upworthy used clickbait headlines and differs from traditional news sources; the conclusions are at the level of news stories, not individuals; and they say nothing about readers' subjective experience. Headlines received fewer clicks when published later in the site's history, which the authors read as suggesting that clickbait was most effective when it was new to readers. Moreover, what the study measures is consumption, which they describe as private behavior, unlike sharing, which is public and influenced by social factors such as signaling group identity and maintaining reputation; this may explain why fear and anger weigh more in the decision to share. The explanation is the authors' and was not tested on the article's data.

The other two studies also appear in Chapter 9 and come in here only for comparison. Rathje et al. (2021) counted, in 2,730,215 posts by news outlets and members of the U.S. Congress on Facebook and Twitter, the words referring to the opposing political side (the out-group) and to one's own side (the in-group), and related these counts to the number of shares. Posts referring to the out-group were shared about twice as often, and each out-group term was associated, on average, with an increase of about 67% (which the authors describe as an increase in the odds of sharing), an effect 4.8 times that of negative affect language and 6.7 times that of moral-emotional language (Section 9.3). Brady et al. (2021) studied the producer's own expression of outrage. In 12.7 million tweets from 7,331 users, a 100% increase in social feedback (likes and shares) to the previous day's outrage was associated with an expected 2 to 3% increase in the next day's outrage; in two experiments with 120 participants, who chose tweets to post in a simulated network, the choice of outrage tweets increased with feedback and followed the norm displayed in the feed (Section 9.5).

The comparison teaches more by what it rules out than by what it allows. "Negative content is more reinforced" is not a proposition describing a single process; there are at least three: headlines with negative words are clicked more (Robertson et al.), posts mentioning the out-group are shared more (Rathje et al.), and the expression of outrage tracks the feedback received (Brady et al.). These are distinct responses, by people in distinct positions in the cycle that will be seen in Chapter 9, and nothing requires that the same consequence maintain them. The pattern of discrete emotions shows the difference: anger, which the popular idea holds to be the engine of virality, had no significant effect on clicks, whereas out-group language predicted "angry"

reactions to Facebook posts (Rathje et al.). The authors' explanation that consuming is private and sharing is public is a functional hypothesis, and the chapter takes it under that name.

TABLE 8.1

Three responses to emotional and moralized content: what each study measured and what it does not allow us to say

Study	Response and material	Content property	What they observed	Caveat
Robertson et al. (2023)	Clicking on the headline (consumption, private behavior). 12,448 Upworthy randomized tests; 53,699 headlines; more than 205 million impressions	Negative and positive words and discrete emotions, by dictionary	Each negative word: +2.3% clicks (headline of average length); positive words: about –1.0%; sadness increased, joy decreased; anger had no significant effect	Clickbait outlet; level of the story, not the individual; nothing about the reader's experience
Rathje et al. (2021)	Sharing and reacting (public behavior). 2,730,215 posts by U.S. news outlets and members of Congress, on Facebook and Twitter	Reference to the political out-group; negative affect and moral-emotional language	Out-group: about twice as many shares; +67% per term, 4.8 and 6.7 times the effect of negative affect and moral-emotional language; "angry" the most popular reaction in seven of eight datasets	Correlational; period before the 2020 election; unclear how algorithmic choices contribute
Brady et al. (2021)	Posting outrage (producer behavior). 12.7 million tweets from 7,331 users; two experiments with 120 participants	Expressed moral outrage, measured by a classifier	+2 to 3% outrage on the next day for each +100% feedback; in the experiments, the choice of outrage increased with feedback and followed the norm	Observational studies do not allow causal inference; users identified through episodes of public outrage; measures text, not emotion

Note. The figures come from the articles. The page numbers for Robertson et al. (2023) are those of Nature Human Behaviour; Rathje et al. (2021) and Brady et al. (2021) have no journal pagination. The "Content property" column describes form (words, reference to a group, a classified expression); the function the form serves for whoever clicks, shares, or posts was not measured in any of the three.

There is also the difference between form and function. The properties measured are properties of form: words from a dictionary, reference to a group, an expression of outrage classified by a semantic model. Operant analysis asks about function, that is, what makes a negative word in a headline an occasion for clicking, or a mention of the out-group an occasion for passing something on. Two answers appear in the authors, signaling group identity and maintaining reputation (Robertson et al.) and satisfying social identity motives (Rathje et al.), and both are proposed functions that no study manipulates. Without function, "negative content" is topography, and the functional treatment presented in Chapter 3 shows that the same form often serves distinct functions or is under multiple control.

DATA	In randomized tests with more than 2.7 million clicks, negative words increased the click-through rate and positive words decreased it (Robertson et al., 2023); posts about the political out-group are shared roughly twice as much as those about the in-group (Rathje et al., 2021); the next day's outrage tracks the feedback received by the previous day's, with a small effect (Brady et al., 2021).
INTERPRETATION	Formal properties of content (words, reference to the out-group) are associated with more clicks, shares, and reactions, and the association differs according to the response measured, which suggests distinct operants with distinct consequences.
HYPOTHESIS	What makes moralized content an occasion for passing it on is the function it serves for the person who passes it on, for example signaling group identity or maintaining reputation (the authors' reading). None of the studies manipulates function.

8.3 The algorithm as programmer: how large an audience each emission receives

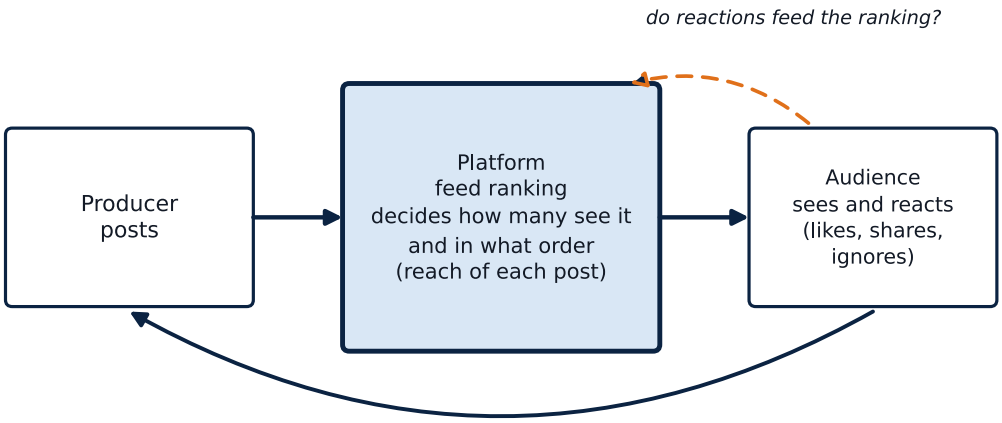
In the previous sections, the producer posts and the audience reacts; what is missing is the term that decides how many people see each emission and in what order, and that is the subject of this section. On a platform with a ranked feed, this term is a set of machine learning models; in operant terms, it is the rule that sets, for each emission, the size of the audience that will be able to react to it. Skinner (1957) already noted that the weakening effects of intermittent or delayed reinforcement are, in verbal behavior, partly offset by the fact that the effects of verbal behavior can be multiplied when the same sound

waves reach many ears or the same page reaches many eyes. The platform controls this multiplication: ranking decides how many eyes will reach the emission and react to it.

Figure 8.1 is a model, and the note says where there are data. The solid arrows between producer, platform, and audience are supported by the audit by Huszár et al. (2022), described below, and the one that carries the counts back to the producer by the studies in Chapter 9. The dashed arrow, in which reactions feed the ranking, is a hypothesis. Brady et al. (2021) state that feed algorithms, by determining how many users see a post, affect social feedback and may influence moral behaviors by exploiting reinforcement learning tendencies; this is the authors' interpretation, and the study does not measure the algorithm.

FIGURE 8.1

The algorithm as programmer of the feed's contingencies



Solid line: relation with data (reach under ranking and under chronological order, Huszár et al., 2022; posters' sensitivity to feedback, Lindström et al., 2021). Dashed: hypothesis; what the ranking uses as input is not known from the chapter's sources.

Note. Original diagram by the author, without data. Solid line: relation with data (reach under ranking and under chronological order, Huszár et al., 2022; posters' sensitivity to feedback, Lindström et al., 2021). Dashed: hypothesis; what the ranking uses as input is not known from the chapter's sources.

Huszár et al. (2022) are the verified source that compares personalized ranking with chronological ordering in their effect on the audience. In 2016, Twitter introduced machine learning algorithms to order the Home timeline according to a personalized relevance model; before that, it showed tweets from followed accounts in reverse chronological order. When it introduced person-

alization, it excluded from it a control group of 1% of all users worldwide, who kept the chronological timeline; the treatment group was a sample of 4% of the remaining accounts. On June 5, 2020, the experiment included 58,087,969 account identifiers, 5% of all accounts worldwide, 20% in control and 80% in treatment. Amplification is the ratio between the reach of a set of tweets in the treatment group and in the control group: 0% means equal reach, and values above 200% indicate an audience three times that of the chronological timeline. Reach was counted in linger impressions (Section 8.1).

All the parties analyzed had positive amplification from personalization, in some cases above 200%. Comparing the largest right-wing or center-right party with the largest left-wing or center-left party in each country, with the exception of Germany, the former had significantly greater amplification; the effect was strongest in Canada (Liberals 43%, Conservatives 167%) and in the United Kingdom (Labour 112%, Conservatives 176%). The seven countries are the United States, Japan, the United Kingdom, France, Spain, Canada, and Germany; Brazil is not among them.

The result changes in nature when the unit is the politician rather than the party. At the individual level, amplification varied widely within each party, from about 400% to below 0%, and a permutation test (which compares the result with the one obtained by randomly shuffling parties among politicians) detected no statistically significant association between party and each politician's amplification. The authors see no contradiction: different politicians may reach overlapping audiences, so group-level correlations can emerge even without individual-level correlation. In countries with well-represented far-left or far-right parties, the amplification of these parties was generally lower than that of moderate parties, and the analysis does not support the hypothesis that personalization amplifies extreme ideologies more than moderate voices; the authors note that strong partisan bias in news reporting was associated with higher amplification, which does not imply the promotion of extreme ideology. Nor was there a consistent pattern between government and opposition: in the United Kingdom amplification favored the governing Conservatives, and in Canada the opposition Conservatives.

For news, the analysis used 100,575,284 tweets pointing to 6,258,032 articles, classified by AllSides, a database that labels the political position of each outlet on a five-point scale from partisan left to partisan right: more partisan sources were more amplified than centrist ones, and the partisan right slightly more than the partisan left. With Ad Fontes, another independent labeling database, the partisan left had low amplification (10.5%), and the authors say the result depends on the political label database.

The limitations are stated. The control group is not insulated from indirect effects, because those in it see content shared by those in treatment; for this reason the experiment does not satisfy the assumption known as SUTVA (that the outcome of one account does not depend on the treatment of others), which precludes unbiased estimates of causal effects, and the statistics probably underestimate them. In addition, the treatment, that is, the ranking experience, changed over time. As for the mechanism, the authors say that understanding it invites further study and cite, as a possible explanation, that different parties may pursue different strategies on Twitter. Finally, the control group was not created for research but to improve the algorithm and provide a business baseline; the study went through review by Twitter's legal and privacy teams, not by an ethics committee.

Mosleh and Rand (2022), described in Chapter 1, approach the same problem from the audience side. They built falsity scores for 816 U.S. public elites from PolitiFact fact-checks: each fact-check receives a veracity rating (from 1, true, to 0, "pants on fire"), and the mean of that public figure's ratings, subtracted from 1, gives the falsity score. They also built, for each Twitter user, an exposure score based on whom the user follows. Exposure to elite misinformation was negatively correlated with the quality of the news the users themselves share and positively correlated with estimated conservative ideology. The authors say the approach is based on the network of followed accounts and does not capture exposure beyond them, such as through retweets or algorithmic recommendations, and suggest that the score could be used in future work to distinguish the role of algorithmic recommendations from that of individual preferences. The falsity scores come from a single fact-checker and depend on what it checked, a selection issue that Chapter 2 addressed.

What the audit measures is the difference between two programmers, personalized ranking and chronological order, not what the ranking optimizes nor whether reactions enter the calculation. For the chapter's argument, this suffices for a limited conclusion: the platform is a variable that alters the reach of each emission, and the alteration is not uniform across emitters, so the analysis of the producer needs two variables, what the audience does with what it receives and how much it receives, and the second is set by the programmer. Whether the difference between groups reflects platform preference, party strategy, or audience behavior is what the data do not decide; the same criterion is applied to all sides: amplification is a datum about reach, not about bias or intention.

DATA	In six of seven countries, the largest right-wing or center-right party had significantly greater amplification from personalization than the largest left-wing or center-left party; in Germany the difference was not significant; at the level of individual politicians, there was no association between party and amplification; the result for news depends on the label database (Huszár et al., 2022).
INTERPRETATION	The platform is a variable that determines the reach of each emission unequally across producers, and reach enters into what the producer can receive (Skinner, 1957). The difference between the two orderings does not reveal what the ranking optimizes.
HYPOTHESIS	If the reactions received feed back into the ranking, the platform closes the loop of Figure 8.1: what received reactions comes to be shown to more people, and the producer receives more of the same. No source in the chapter shows what the ranking uses as input.

8.4 Intermittent reinforcement and engagement

The most repeated analogy about social media is the slot machine: the user pulls the lever, and the prize comes now and then. This section states where the analogy's vocabulary comes from, what the experimental literature describes for each schedule, and what the verified sources support about the feed, which is little.

CONCEPT IN USE

Intermittent reinforcement (Skinner, 1957)

Intermittent reinforcement is reinforcement in which only some of the responses, or of the occasions, are followed by the reinforcer. For Skinner (1957), reinforcement that is merely intermittent, like the occasional lucky throw of the dice, is enough to keep a response strong; for this he refers to the experimental study of nonverbal behavior. In verbal behavior, intermittency is the rule: the effect depends on the presence and activity of a reinforcing organism, the listener, whose behavior is neither inevitable nor generally predictable. The named schedules (ratio and interval, fixed and variable), from which the slot-machine analogy borrows its vocabulary, come from the experimental literature summarized in the next box.

Skinner's argument has two consequences for the chapter. The first is that, for the speaker, intermittency is the normal condition, not a peculiarity of social media: the listener may not be there, may not react, and the effect on the speaker is intermittent by construction. The second is that what the platform changes is not the presence of intermittency but its form: how many listeners can react (Section 8.3) and the visibility of the count to the producer and to other listeners. This is the reading proposed here.

CONCEPT IN USE

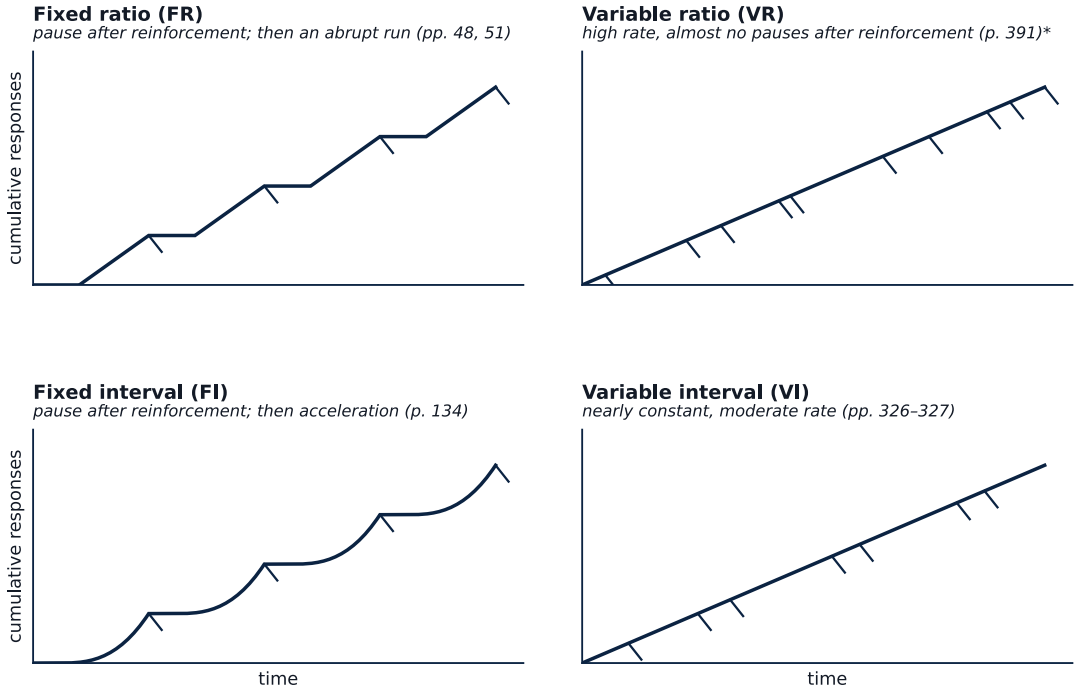
Schedules of reinforcement (Ferster and Skinner, 1957)

A schedule of reinforcement can be specified without referring to what it does to behavior: the reinforcer may depend on the time elapsed since the preceding reinforcement or on the number of responses emitted since then, and the schedule may be fixed or variable, which yields four basic schedules. In fixed ratio (FR), reinforcement follows the completion of a set number of responses, counted since the last reinforcement; in variable ratio (VR), the required number varies at random around a given mean; in fixed interval (FI), the first response after a set interval, timed from the last reinforcement, is reinforced; in variable interval (VI), the interval varies at random around a given mean. The authors distinguish the function of shaping the form of the response and stimulus control from the function of maintaining an active repertoire, and they say that most reinforcements are intermittent. Most of the experiments use pigeons pecking a key to obtain food, and some use rats pressing a lever. At the human level, they say that schedule analysis had already shown its usefulness in research on psychotic behavior and in classroom teaching techniques, and that further applications, in law and penology, religion, industry, and commerce, hold considerable promise.

What each schedule generates in the authors' experiments is the term of comparison the analogy needs. In fixed ratio, stabilized after prolonged exposure, pauses after reinforcement are occasional and, when they occur, the shift to the terminal rate is abrupt (FR 60, after more than 4,000 reinforcements); with larger ratios the pauses grow longer, up to 10 minutes on FR 120. In fixed interval, the normal stable state has a pause after each reinforcement, followed by acceleration to a terminal rate that is usually moderate. Variable interval is arranged to generate a steady rate, by preventing any aspect of the bird's behavior from acquiring discriminative properties, and most VI schedules include an interval of zero or near zero to prevent a pause after reinforcement; in a bird on VI 1 minute, after 16 sessions, the rate was approximately constant, with no change in rate tied to any feature of the schedule. Variable ratio, like fixed ratio, arranges for differential reinforcement of high rates, and there are very few instances of reinforcement after a pause. Figure 8.2 traces these patterns in a diagram.

FIGURE 8.2

Response patterns that Ferster and Skinner (1957) describe for the four basic schedules: conceptual diagram



Original diagram, no data: slopes and durations are illustrative, the vertical scales are independent (do not compare heights across panels), and the oblique marks indicate reinforcements. It traces what Ferster and Skinner (1957) describe in text, for pigeons pecking a key to obtain food.
** With large means, variable ratio produces variety, with long pauses and bursts (pp. 394–397).*

Note. Original diagram by the author, without data, based on the textual descriptions in Ferster and Skinner (1957), whose records are from pigeons. The cumulative record, in which responses add up over time, is the authors' basic datum. Slopes and durations are illustrative, and the vertical scale of each panel is independent.

Variable ratio, the schedule the analogy invokes, is also the one for which the authors record the greatest variety: it produces diverse performances depending on the distribution of the numbers of responses required, and the bird stops responding if the mean goes beyond a certain value. With a mean of 360 responses, one bird sustained in one session a rate of about 7 responses per second and, in the next, a burst right after reinforcement followed by long pauses, one of more than 1 hour; the authors treat the two performances as the limits of a stable state. The other bird never sustained a high rate and responded in bursts of 10 to 75 responses, separated by pauses of 10 to 70 seconds.

Gunnarsson et al. (2026) write, in their discussion, that their findings illustrate how platforms leverage variable-ratio schedules: since likes are never certain, every successful episode of attention ends up being disproportion-

ately reinforcing. This is the authors' interpretation; the schedule was neither manipulated nor measured, and it could not have been in a hypothetical purchase task. The argument also jumps from intermittency to variable ratio: that likes are not guaranteed shows that reinforcement is intermittent, and calling it variable ratio further requires that delivery depend on the number of responses and vary around a mean (Ferster and Skinner, 1957). The authors further suggest that those who use social media mainly to obtain attention-related reinforcement may be especially prone to problematic usage patterns, on the basis of the interaction between the social media use scale and the function condition for consumption at minimal cost.

Two studies from Chapter 9 are compatible with the hypothesis without testing it. Lindström et al. (2021) show, with data from Instagram and online forums, that posting rate tracks the rate of likes according to a hyperbolic function, which rises quickly and then flattens (compatible with the quantitative law of effect), better than according to a linear relation. Ceylan et al. (2023) rewarded with points, over 80 trials, the choice of sharing or not sharing headlines according to a veracity criterion and show that the sharing pattern installed by this feedback persists in the test in which feedback is withdrawn. In neither study was intermittency the manipulated variable: persistence without feedback is compatible with a history of intermittent reinforcement and with other histories, and the design does not separate them. Lindström et al. deal with when people post, not with what they post, and note that there are many motives for posting besides reward (Section 9.2).

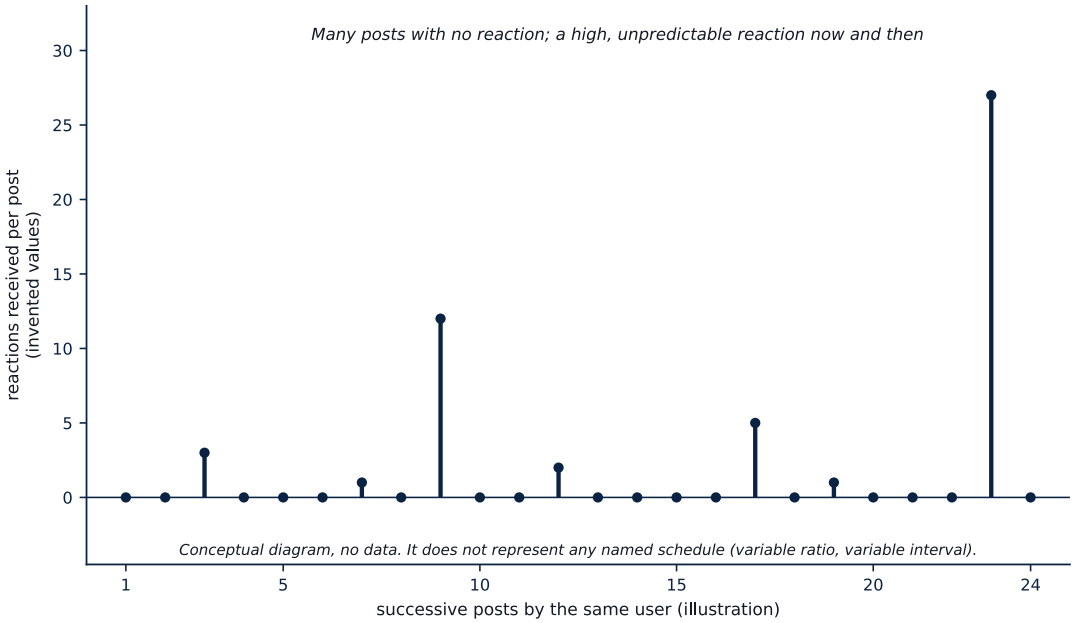
Calling the feed a variable-ratio schedule requires at least three things. The first is the distribution of reactions per post, for the same producer, over many posts: variable ratio presupposes a series of ratios with a given mean, not merely a fluctuating count. The second is the relation between the number of responses, posting or checking reactions, and the delivery of the reinforcer: in ratio schedules what counts is the number of responses since the preceding reinforcement, and in interval schedules, time. The third is the response pattern the schedule generates, compared with that of Figure 8.2 and with a baseline, for example the time until the next post after one that received many reactions and after one that was ignored. No source in the chapter provides the first or the second, and the third exists, for the feed, only as a question. Figure 8.3 serves to show the loose points of the analogy, with invented values.

Three loose points appear in the figure. The first is that the reaction to a post is not a discrete reinforcer that either comes or does not, like a slot machine's token or the operation of the food magazine in the authors' experiments: it is a count, which grows over time and depends on how many people see the post

(Section 8.3). The second is that there are at least two candidate operants, posting and checking reactions, and they may be under different delivery rules: the reinforcer for checking is available as long as someone reacts, and the reinforcer for posting depends on the post. The authors studied, in experiments they describe as exploratory, two responses with their own concurrent schedules, and the ratio response dominated the interval response; whether something similar occurs between posting and checking is also a question, not a result. The third is the most important for this analysis: if what determines how much each post receives is partly its form, and not chance, the contingency is differential, since it selects forms rather than merely dosing frequency, which would suggest at least that different schedules are superimposed or competing with one another. A schedule of reinforcement describes how much and when; the selection of forms describes what, and the authors distinguish shaping the form of the response from maintaining the repertoire. The selection of forms is what matters for this chapter's question, and it is what the chapter's sources best support, in Section 8.2. All of this is interpretation.

FIGURE 8.3

Reactions received by successive posts of the same user: conceptual diagram



Note. Original diagram by the author. The values were invented to show the idea of intermittent reinforcement and do not represent data, nor a named schedule (variable ratio, variable interval).

DATA	For Skinner (1957), intermittency is the rule of verbal behavior and is sufficient to maintain a response; Ferster and Skinner (1957) define the four basic schedules and describe, in pigeons, a pattern for each (pause and run in fixed ratio, pause and acceleration in fixed interval, nearly constant rate in variable interval, high rate with rare pauses in variable ratio, with variety when the mean is large); Gunnarsson et al. (2026) attribute variable-ratio schedules to platforms without measuring them; posting rate tracks the rate of likes according to a hyperbolic function (Lindström et al., 2021).
INTERPRETATION	Intermittency is a general condition of verbal behavior. That the delivery of reactions to those who post is intermittent is plausible and consistent with the data, but it is a premise, not a result; and intermittency is not, in itself, variable ratio.
HYPOTHESIS	Delivery follows a named schedule (variable ratio or variable interval), and the usage pattern results from it. Testing this would require the distribution of reactions per post, the relation between responses and delivery, and the response pattern compared with that of Figure 8.2; no source in the chapter provides the first two, and the third has not been examined for the feed.

8.5 Experiments on the algorithm itself: what changed and what did not

Experiments in which the delivery rule was manipulated are the most direct evidence of what the platform does. The verified sources include two, that of Huszár et al. (2022), described in Section 8.3, and that of Nyhan et al. (2023), in addition to the publisher's test in Robertson et al. (2023), which manipulates another point in the chain. Table 8.2 sets them side by side.

The experiment by Nyhan et al. (2023) is described in Section 6.2, and what matters here is what it manipulated. For participants assigned to treatment, all content from friends, groups, and Pages predicted to share the participant's political leaning was downranked in the feed, from September 24 to December 23, 2020; the control group's feed was not systematically altered. According to the authors, the intervention aimed solely at reducing exposure to like-minded sources, with the largest demotion strength that a pre-test indicated would not leave the feed nearly empty. Exposure fell from 53.7% to

36.2%, and there was no measurable effect on the eight preregistered attitude measures; those whose exposure was reduced engaged more, per exposure, with the like-minded content they still saw.

TABLE 8.2

Three manipulations along the producer–platform–audience chain: what changed and what did not

Study	What was manipulated	What they measured	What changed	What did not change or was not measured
Robertson et al. (2023)	Headline variations of the same story, randomly assigned to Upworthy readers (manipulation by the publisher, not the platform)	Click-through rate of each variation	Negative words increased clicks (2.3% per word, in a headline of average length); positive words decreased them (about 1.0%)	What the publisher went on to produce; the individual; the reader's experience
Huszár et al. (2022)	Feed order: personalized ranking (treatment, 4% of accounts) versus chronological order (control, 1%)	Reach, in linger impressions, of tweets by legislators from seven countries and of news links	Greater reach with personalization for all parties; greater for the largest right-wing or center-right party in six of the seven countries	Association between party and amplification, by politician; causal mechanism; indirect effects on the control group; what producers did afterward
Nyhan et al. (2023)	Downranking, in the Facebook feed, of content from like-minded sources (09/24 to 12/23/2020; n = 23,377)	Share of content from like-minded and cross-cutting sources; eight preregistered attitude measures	Exposure to like-minded sources from 53.7% to 36.2%, and to cross-cutting sources from 20.7% to 27.9%; more engagement, per exposure, with the remaining congenial content	Attitudes (effects of ±0.12 standard deviation or more ruled out); effects of prior use and effects accumulated over years; the producer

Note. The figures come from the articles. The page numbers for Robertson et al. (2023) are those of Nature Human Behaviour and those for Nyhan et al. (2023), of Nature; Huszár et al. (2022) was read without print pagination. "Changed" and "did not change" refer to what each study measured.

Section 6.1 left a question about the boundary of the bubble: how much of what comes from outside the verbal community reaches the speaker. The delivery rule is one of the variables that decide it, and the experiment shows that the rule moves it: with the downranking of congenial content, exposure to cross-cutting sources rose from 20.7% to 27.9%, and exposure to sources neither congenial nor cross-cutting, from 25.6% to 35.9%. What the experiment does not measure is what the figure in Section 6.1 asked: whether correction directed at what the speaker said came to reach the speaker. Exposure to cross-cutting sources is a share of the feed, not a correction of the statement.

What the two experiments have in common is what they do not measure. Both intervene in delivery and measure what the listener receives or reports: the reach of politicians' tweets in one case; users' exposure and attitudes in the other. Neither measures what the producer did after seeing their reach change. The ranking in Huszár et al. altered the audience size of each emission, and the downranking in Nyhan et al. reduced the reach of like-minded content; for the producer, both are changes in the reach contingency. The question from which the chapter started, whether and how those who post adjust what they post to this rule, could be answered only with series of producers' posts before and after changes in the delivery rule. The platforms have these records; the verified studies do not provide them, which limits access to the possible differential attentional functions involved.

None of this warrants concluding that the delivery rule does not matter (Nyhan et al. found no effect on reported attitudes, after three months and on a single service) or that it explains what is said (Huszár et al. measured reach, not what is said). The null effect on attitudes has, moreover, explanations that the authors themselves list: congenial political content is a fraction of what people see, Facebook is a fraction of the information diet, persuasion is difficult, and the design does not capture prior use or effects accumulated over years.

DATA	Personalized ranking and chronological order differ in the reach of parties and news sources (Huszár et al., 2022); reducing exposure to like-minded content by one third did not change eight attitude measures over three months (Nyhan et al., 2023).
INTERPRETATION	The delivery rule alters reach, and reach is a variable on which what the producer can receive depends. In both experiments the dependent variable belongs to the listener or to reach, not to the producer.

HYPOTHESIS Producers adjust what they post to changes in reach produced by the delivery rule. Testing this would require measuring production before and after a rule change; no verified source does so.

8.6 Rival interpretations and limits

The chapter's reading, that the attention distributed by the algorithm functions as a reinforcer for those who post, has rivals. The box below brings them together, one by one, and the chapter's limits come after it.

Rival interpretations of the same body of data

The user chooses, and the algorithm reflects. What the audience reacts to and what the algorithm shows are confounded: Mosleh and Rand (2022) say their score does not capture exposure through algorithmic recommendation and suggest it be used to separate the role of recommendations from that of preferences; Nyhan et al. (2023) note that the algorithm changes tested do not cancel users' propensity to seek congenial information; and Rathje et al. (2021) say they do not know how algorithmic choices contribute.

Attention is not the only reinforcer, nor the main one. In the task of Gunnarsson et al. (2026), escape, information, and automatic positive reinforcement produced more demand than the search for likes, and the authors point out that many people use social media for multiple, overlapping functions.

Norm and identity instead of reinforcement. What is observed may be conformity to network norms (Brady et al., 2021) or expression of group identity (Rathje et al., 2021); see Section 9.8.

Novelty and clickbait. The effect of negative words may be specific to a clickbait outlet and to a phase in which the device was new to readers (Robertson et al., 2023).

Small effects. The effects are 2 to 3% (Brady et al., 2021) and about 2.3% per word (Robertson et al., 2023); they may scale up, but that is an argument, not a result.

Reach is not bias or intention. Amplification by party or source is a datum about reach, depends on the label database, and has no identified mechanism (Huszár et al., 2022).

Data only the platform has. Twitter's control group existed for business purposes and was reviewed by legal and privacy teams, not by an ethics committee; the treatment changed over time (Huszár et al., 2022). Outsiders without access to the records cannot replicate or extend the study.

Four limits apply to the whole chapter. First, no source is Brazilian, and the platform sources deal with Twitter and Facebook, in the United States and, in the case of Huszár et al. (2022), in six other countries, among which Brazil is not included; platforms change (the treatment in the Twitter experiment changed over time), and nothing here speaks to WhatsApp, YouTube, or TikTok. Second, the broadest datum on the delivery rule (Huszár et al., 2022) measures reach, not producer behavior. Third, the data on moralized content are mostly associational, and the effects are small. Fourth, the chapter does not show that the feed is a schedule of reinforcement: it shows that, for verbal behavior, intermittency is the rule and that the platform's delivery rule alters reach; the rest is a hypothesis that calls for data only the platforms have.

For those who want to use the chapter in the analysis of utterances (Chapters 3, 5, and 7), the consequence is practical. The reach of an utterance is a variable that the behavior analyst does not control and that alters what the utterance can receive; comparing one politician's utterance with another's without recording the reach of each, and what the platform did with it, confounds two variables. When reach is not known, the analyst should say so.

Summary

The listener's attention is, in Skinner's (1957) analysis, a generalized reinforcer that strengthens verbal behavior without respect to form, and on platforms what reaches the producer is a count of reactions that approximates it. The hypothesis that this count is the dominant reinforcer for those who post has partial support in the studies of Chapter 9 and a caution in that of Gunnarsson et al. (2026), in which demand for using social media to obtain attention was low (Section 8.1). The data on emotional and moralized content do not form a single finding: negative words increase clicks, reference to the outgroup increases shares, and outrage tracks the feedback received; these are distinct responses, measured by properties of form, and the function that maintains them was not measured (8.2).

The platform is the third term. Ranking sets the reach of each emission, and the Twitter audit shows that it differs from chronological ordering, with greater amplification for the largest right-wing or center-right party in six of seven countries, no association at the level of individual politicians, and results for news that depend on the label database; the mechanism is unknown (8.3). The hypothesis that delivery follows a variable-ratio schedule is consistent with Skinner, who takes intermittency to be the rule of verbal behavior, but no source measures it; the schedules literature (Ferster and Skinner, 1957) gives, for pigeons, one pattern per schedule, with variety in variable ratio with large

means, and the comparison with the feed calls for data only the platforms have. The analogy also has loose points: the reaction is a count, there are two candidate operants, and, if the form of the post determines what it receives, what we have is selection of forms (8.4). The two experiments on the algorithm changed reach and exposure and did not change reported attitudes; neither measured what the producer did afterward (8.5).

Chapters 3, 5, and 7 take from this chapter the idea that reach is a variable to be recorded in the analysis of any utterance; Chapter 3 also takes the distinction between function and form that the content properties leave open. Chapter 6 receives a partial answer to the question of how far the delivery rule makes the boundary of the bubble permeable: the rule moves it, and what crosses the boundary as correction was not measured. Chapter 11 (Section 11.2), in dealing with measures of diffusion, returns to the distinction between reach, exposure, and reaction. Chapter 12 (Section 12.6) starts from what did and did not change with the delivery rule to discuss interventions on platform contingencies.

References

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), Article eabe5641. https://doi.org/10.1126/sciadv.abe5641

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. Appleton-Century-Crofts. https://doi.org/10.1037/10627-000

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Gunnarsson, K. F., Rzeszutek, M. J., Malkin, A., Tewogbola, P. O., Han, S., & Jacobs, E. A. (2026). A function-based examination of social media demand. *Behavior Analysis in Practice*. Advance online publication. https://doi.org/10.1007/s40617-026-01197-7

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), Article e2025334119. https://doi.org/10.1073/pnas.2025334119

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), Article 1311. https://doi.org/10.1038/s41467-020-19607-x

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), Article 7144. https://doi.org/10.1038/s41467-022-34769-6

Nyhan, B., Settle, J., Thorson, E., . . . Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137–144. https://doi.org/10.1038/s41586-023-06297-w

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822. https://doi.org/10.1038/s41562-023-01538-4

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

instituto WALDEN4 | ACADEMIA WALDEN4



EDUCATIONAL MATERIALS

Where to start reading behavior analysis

Scientific articles, journals, and reference institutions gathered for training in behavior analysis and evidence-based practice.

Free access, no sign-up required

walden4.com.br/materiais

CHAPTER

9

The Producer–Audience Cycle

In 1956, Hildum and Brown published an experiment in which an interviewer telephoned 40 male students at the Harvard summer school and read them 15 statements about the university's General Education philosophy, each with four alternatives, from “strongly agree” to “strongly disagree.” Half of them heard “Good,” in a neutral tone, after some of their answers; the other half heard “Mm-hmm” at the same points. Within each half, ten students heard the reaction after answers favorable to General Education and ten after unfavorable ones (Hildum & Brown, 1956). With “Good,” the mean attitude score (lower is more favorable) was 24.95 in the group that received the reaction after favorable answers and 31.75 in the group that received it after unfavorable ones. Of the 20 students who heard “Good,” eight noticed that the interviewer was saying the word, and only one guessed that it signaled approval; all of them denied that their answers had been influenced by him.

The authors' question was one of method: if an interviewer can influence the respondent unintentionally, the influence operates outside the awareness of everyone involved, and the opinion survey becomes biased. This chapter asks the question from the speaker's side. What does the listener's reaction do to what is said, and what does what is said do to the listener? In 1956 the reaction was a word from an interviewer. Today it is a public count of likes and shares, distributed by an algorithm. The chapter first gathers the older experimental evidence (Section 9.1) and then the platform evidence (9.2, 9.3, 9.5, and 9.6), proposes a scheme that unites them (9.4) and a hypothesis at the cultural level (9.7), and states at each step what is data, what is interpretation, and what is hypothesis.

9.1 Opinions can be reinforced: the experiments of 1956 to 1969

The Hildum and Brown experiment that opens the chapter has predecessors and successors. The verbal reinforcement studies that preceded it took a word or a class of words as the response unit. In one of them, the experimenter murmured "Mm-hmm" whenever the student said a plural noun; in others, the experimenter said "Good" after sentences that began with the chosen pronoun; in all of them, the subjects said they had not noticed any relation between what they said and the reaction (Hildum & Brown, 1956). The authors argue that an opinion interviewer would react to categories of content rather than to specific words, and they used the telephone so that the vocal reaction would be the only one possible.

Scott (1957) created the approval situation in the classroom. Students in 29 general psychology sections wrote down what they thought about three controversial campus issues (universal military training, nighttime curfew hours for women students, and the emphasis on football), and each section was given one of them for debate. In each section, the instructor chose two people who were clearly in favor and two who were clearly against and invited them to an "experiment" on how much they could influence the class, on the condition that each argue the side opposite to his or her own. Two weeks later, the pairs debated in front of the class, with three minutes of presentation and two of rebuttal for each side, and the class voted by ballot on which debater had done better. The vote was staged: the instructor pretended to count the ballots, and who would "win" in each pair had been decided in advance, with the winning side alternated. Immediately afterward, everyone rewrote their opinion on the issue, and the texts, copied onto cards identified only by number, were coded on a seven-point scale, from very favorable to very unfavorable. Of the 58 pairs assigned, 36 produced complete data, in 19 sections and with six instructors; another 36 students, with a clear opinion but no debate, formed the control group. The winners changed on average 1.25 points in the direction of the side they had argued, the losers changed 0.17 points in the opposite direction, and the controls 0.31 points in the same direction as the winners. The difference between winners and losers was statistically significant ($t = 3.97$; $p < .001$); t is the test statistic that compares two means, and p is the probability of obtaining a difference as large as this one, or larger, if the groups did not actually differ, so that, by convention, a p below .05 is called "significant." The difference between winners and controls was also significant ($p < .02$), and the difference between losers and controls was not (Scott, 1957).

Spielberger et al. (1962) studied verbal reinforcement in a task that is not about opinion, sentence construction, and asked whether the effect requires that the person be aware of the contingency. Each of 45 female undergraduates, tested individually, received 100 cards, each with a past-tense verb and the pronouns *I*, *we*, *you*, *he*, *she*, and *they* in different orders, and had to compose a sentence that began with one of the pronouns and contained the verb; each sentence was a trial. On the first 20, no one was reinforced, which measures the baseline frequency of sentences beginning with *I* or *we*. The participants were then randomly assigned to three groups of 15, which received neutral instructions, instructions that asked for the first sentence that came to mind, in a hurry, and discouraged the search for a principle, or instructions that said there was a correct way to build the sentences, so that awareness of the contingency would vary. Over the next 80 trials, in each group 10 participants heard "good," said in a neutral tone, after every sentence beginning with *I* or *we*, and five, who formed the control group, heard nothing. At the end, an interview with increasingly specific questions, which avoided suggesting the correct contingency, classified each participant as aware or unaware that "good" followed those sentences. Of the 30 reinforced participants, 19 were classified as aware; only they increased their use of *I* and *we*, and none of the 11 unaware participants increased by more than one response (Spielberger et al., 1962).

Insko (1965) took the procedure to the telephone and measured whether the effect lasted. Six graduate assistants each telephoned 12 students chosen at random from the University of Hawaii's introductory psychology course (72 in all) and read 14 opinion statements about creating a spring Aloha Week, such as that it would bring in more tourist money or more traffic congestion; the statements did not say whether the idea was good or bad. The student answered whether he or she agreed or disagreed with each one, and each assistant said "good" after positive-attitude answers for six of the students and after negative-attitude answers for the other six. Each answer was worth 0 to 3 points, and the mean total was 28.22 among those reinforced for positive answers and 20.69 among those reinforced for negative ones. About a week later, the experimenter, who was the course instructor, handed out to the 225 students a questionnaire on local issues, with the Aloha Week item two thirds of the way through. Of the 72 contacted by telephone, 70 responded. The group previously reinforced for positive answers averaged 10.40, in the region of the scale labeled "slightly in favor," and the group previously reinforced for negative answers averaged 8.20, in the region labeled "uncertain" ($t = 2.03$; $p < .05$) (Insko, 1965).

Insko and Melson (1969) used the same procedure to ask whether the effect depends on the person knowing that he or she is in an experiment, the demand characteristics hypothesis: the participant would act as he or she supposes the experimenter expects. In Experiment I, 72 male students enrolled in introductory psychology at the University of North Carolina were contacted, half through the usual recruitment procedure for laboratory experiments and half by telephone. In both cases, the interviewer said he worked for a survey research center that was polling students' attitudes about pay TV, and those who received the call were not told that they were taking part in an experiment. The participant agreed or disagreed with 12 opinion statements about pay TV, and "good" followed agreeing with statements in favor and disagreeing with statements against (pro condition) or the reverse (con condition); the point in the semester, beginning or end, was also varied. Immediately afterward, everyone answered a questionnaire on awareness of the reinforcement, developed by Dulany in 1962. Experiment II repeated the design with 20 subjects in each of the four combinations of context (laboratory or telephone) and direction of reinforcement, and added a baseline of eight statements without reinforcement. In both experiments, the only significant effect was that of reinforcement (F = 6.10 and 5.62; F is the analysis-of-variance statistic, which compares the variation between groups with the variation within them), with no difference between laboratory and telephone (Insko & Melson, 1969).

Table 9.1 brings together the five studies, with what was done, what was observed, and what the results do not allow us to say. Four of them deal with opinions and attitudes; the Spielberger et al. study uses a task that is not about opinion and is included because it is the main contrary evidence on awareness of the contingency.

TABLE 9.1

Experiments on verbal reinforcement of opinions and attitudes, 1956 to 1969: what was done, what was observed, and what they do not allow us to say

Study	Participants and task	Reaction to what was said	What was observed	Caveat
Hildum and Brown (1956)	40 male students (Harvard summer school), 10 per group; 15 statements about General Education, by telephone	"Good" or "Mm-hmm" after favorable or after unfavorable answers	With "Good," mean attitude score of 24.95 (reaction to favorable answers) versus 31.75 (to unfavorable ones; $t = 3.31$; $p = .01$); with "Mm-hmm," a nonsignificant difference in the opposite direction; replication with 25 students, same pattern	8 of 20 noticed the "Good"; all denied being influenced; of the 24 who ventured a guess, all but one said the interviewer favored General Education
Scott (1957)	72 debaters (36 pairs) from 19 psychology sections, arguing the position opposite to their own; 36 controls	Class vote, decided in advance, declaring the debater winner or loser	Winners changed on average 1.25 points (7-point scale) in the direction of the debate; losers, 0.17 in the opposite direction; controls, 0.31 (winners × losers: $t = 3.97$; $p < .001$)	No debate-without-vote condition; posttest within five minutes; only seven winners reversed their opinion; persistence not assessed
Spielberger et al. (1962)	45 female undergraduates; sentence construction with pronouns (a task not about opinion); 30 reinforced and 15 controls	"Good" after sentences beginning with I or we	Only the 19 subjects who verbalized the contingency increased their responding; none of the 11 unaware subjects rose by more than one response	The authors note the role of the task and the population; attitude toward the reinforcement modulated the effect among the aware subjects

Study	Participants and task	Reaction to what was said	What was observed	Caveat
Insko (1965)	72 students (Hawaii); 14 statements about a spring Aloha Week, by telephone; classroom questionnaire about a week later (N = 70)	"Good" after agreeing or after disagreeing	Telephone: 28.22 (reaction to positive attitude) versus 20.69 (to negative); questionnaire: 10.40 versus 8.20 (t = 2.03; p < .05)	The author speaks of response generality; the experimenter was the instructor of the class that answered the questionnaire
Insko and Melson (1969)	Exp. I: 72 men; Exp. II: 80 (20 per cell); telephone or laboratory; 12 statements about pay TV	"Good" after pro or con answers	Significant reinforcement effect (F = 6.10 and 5.62), equal in the laboratory and by telephone; among those unaware of the contingency, nonsignificant difference in each experiment; combined, 6 of 6 comparisons in the predicted direction (p ≈ .016)	Effect less robust than in an earlier study with the same statements (F = 23.09); only 14 of 152 reported any intention

Note. The figures come from the articles (Hildum & Brown, 1956; Scott, 1957; Spielberger et al., 1962; Insko, 1965; Insko & Melson, 1969). In Hildum and Brown, a lower score is a more favorable attitude; in Insko, a higher one. "Pro" and "con" refer to the topic of the interview.

The table shows a pattern with four parts. First: in the opinion tasks, the listener's approval shifted what was declared in the reinforced direction, in all four studies. Second: the effect appeared outside the situation in which it was produced. Insko (1965) found it a week later, in another medium (a written questionnaire) and in another setting (from home to the classroom), and Insko and Melson (1969) found it to be the same size in the laboratory and by telephone, with subjects who did not know they were in an experiment. Third: the effect is small and fragile. In Scott (1957), only seven of the winners reversed their opinion, the other 14 who changed merely weakened it, and the posttest came no more than five minutes after the vote; in Insko and Melson (1969), the effect was much less robust than in an earlier study with the same statements. And the reinforcer matters: "Mm-hmm" did not work in Hildum and Brown, who suggest that the difference may come from the greater visibility of "Good" (p. 110) and, in part, from its more clearly favorable meaning. Fourth: awareness of the contingency is disputed. In Spielberger et al. (1962), only the subjects who verbalized it learned; in Insko and Melson (1969), among those who did not report being aware of the contingency, the differ-

ence between conditions was not significant in either of the two experiments taken separately, but the six combined comparisons went in the predicted direction, and a reanalysis of an earlier study with the same statements yielded a significant difference. The authors conclude that reinforcement also had an effect without awareness and attribute the difference from sentence construction to the task: giving an opinion is an overlearned response, that is, one that is highly practiced, emitted in few trials and with little novelty.

CONCEPT IN USE

Generalized social reinforcement and audience (Guerin, 1994), revisited from Chapter 3

Chapter 3 (Section 3.3) presented generalized social reinforcement and Guerin's (1994) definition of attitude as a statement maintained by it. Two implications matter for the cycle. The first is that the consequence comes from other people and, in part, does not depend on content, as in the exchange of attitudes that functions as a ritual social event, in which the reinforcement lies in the exchange itself. The second is that the audience is a discriminative stimulus: specific audiences listen and respond to specific attitudes, and so what is declared tends to meet the contingencies of the community that occasioned the declaration. The generalized reinforcer does not operate equally for everyone. Spielberger et al. (1962) point out that verbal conditioning studies assume that "Good" and "Mm-hmm" are generalized conditioned reinforcers, with the same properties for all subjects, and they cite evidence that "Mm-hmm" raised the performance of those who rated it as positive and lowered it in those who rated it as negative (Mandler & Kaplan, 1956, as cited in Spielberger et al., 1962).

Rival interpretations for the shift in opinion under approval

Awareness of the contingency. If the speaker notices that certain answers are approved and starts giving them more often, what we see is rule-governed performance, not an operant shaped without awareness. The results divide by task (Spielberger et al., 1962; Insko & Melson, 1969), and Hildum and Brown (1956) acknowledge that their awareness measures may have been insensitive.

Demand characteristics. If participants know they are in an experiment and act as they suppose is expected, the effect should be larger in the laboratory than outside it. Insko and Melson (1969) found no difference between the laboratory and the telephone without identification as an experiment.

Interview response, not attitude. The shift might hold only for what is said in that interview. Insko (1965) measured a week later and in another medium, and the author himself describes what he showed as response generality.

Contact with opposing arguments. In Scott (1957), winners and losers had the same contact with the opposing arguments, and the difference between them points to the class's approval or disapproval. But the winners' change may be an

effect of reinforcement alone or an effect of contact that reinforcement maintained, and a condition of debaters without a vote, needed to separate the two readings, was missing.

Assumption about the interviewer. In Hildum and Brown (1956), almost everyone who ventured a guess assumed that the interviewer favored General Education, which the authors attribute to assumptions about who researches the topic, not to the pattern of reinforcement.

DATA	In telephone interviews (Hildum & Brown, 1956; Insko, 1965; Insko & Melson, 1969) and in a classroom debate (Scott, 1957), what people declared as their opinion shifted in the direction of the approval they received more than in the opposite or control group, in small effects that vary with the reinforcer and the task.
INTERPRETATION	Declared opinion is sensitive to the social consequences that the audience arranges for what is declared, as Guerin's (1994) analysis predicts. The listener's reaction functions as a consequence of the statement, and the audience as an occasion for it.
HYPOTHESIS	The same sensitivity holds for what is posted when the audience's reaction is a like or a share. None of the five studies uses social media or public, counted reactions; the extrapolation is made here.

9.2 What the producer receives: likes, shares, and reach

On platforms, the one who speaks is a producer: he or she posts and receives back, counted and visible, the public's reactions, the *feedback* of the studies. The studies used here are of two kinds: observational ones, which ask whether the act of posting tracks the feedback received, and experiments in simulated environments, in which the feedback is manipulated.

Lindström et al. (2021) treated posting as free operant behavior, without discrete trials: the equivalent of an animal facing a single lever that can be pressed at any moment, with the like as the consequence. They used four datasets, totaling more than one million posts from more than 4,000 users: Instagram (Study 1, with 2,039 users and, on average, 418 posts per person) and three topical forums, on men's fashion, women's fashion, and gardening (Study 2, with 2,127 users and, on average, 91 posts per person). The forums

were included because on Instagram there are economic motives for posting and the possibility of fraudulent accounts and fake likes, which are less likely in the forums.

The authors carried out two analyses. In the first, they compared two ways of describing how the posting rate varies with the like rate. Animal research has shown that response rate usually grows with reward rate according to a hyperbolic function, which rises quickly and then flattens out, the so-called quantitative law of effect, regarded as a signature of reward-guided behavior. In all four datasets, this function described the data better than a linear relation (mean R^2 of .43 in Study 1 and .37 in Study 2; R^2 is the fraction of the variation in the data that the model explains). In the second, they defined the latency between two successive posts (τPost) and tested whether the history of likes predicted it by Granger causality, which is established when one variable, here likes, improves the prediction of another, here the interval, beyond what the latter's own past values already predict. Previous likes predicted the interval until the next post in all four datasets. Finally, with a learning model, they estimated for each user the average reward rate he or she had been receiving. On Instagram, the interval between posts was about 18% shorter, something like eight hours, when this estimated rate was high than when it was low.

Study 3 was an experiment with 176 participants. During a 25-minute online session, they could post memes whenever and as often as they wished and received from 0 to 19 likes from supposed other users. From the first to the second half of the session, the average like rate changed, from 0–9 per post to 10–19 for some participants and in the opposite direction for others, with the direction counterbalanced. The mean latency to post was about 10.9% longer when the rate was low (0 to 9) than when it was high (10 to 19). The authors caution that there are many motives for posting besides reward, that the study deals with when people post and not with what they post, and that the data do not say whether heavy use is maladaptive.

Ceylan et al. (2023) studied sharing. In four studies (N = 2,476), people with Facebook accounts viewed headlines (in Studies 1 and 2, in Facebook format) and decided, using a simulated share button, whether to share them: there were 16 headlines, eight true and eight false, in Studies 1, 2, and 4. The strength of the habit of sharing news was measured by four self-report items, on a scale from 1 (never) to 7 (always), adapted from a habit index, such as “Sharing news on Facebook is something I do automatically”; the mean of the items gave each person's habit score. In Study 1 (N = 200), on average 32% of the true headlines and 5% of the false ones were shared. Those with the strongest habit (one standard deviation above the mean) shared similar pro-

portions of true (43%) and false (38%) headlines, and those with the weakest (one standard deviation below), 15% and 6%; the 15% most habitual users accounted for 37% of the false headlines shared. In Study 2 (N = 839), participants were randomly assigned either to choose first whether to share the 16 headlines and then judge the accuracy of each one, or to judge the accuracy of all of them before deciding whether to share. Judging accuracy first reduced the sharing of false headlines, but it did not increase discrimination among the most habitual users any more than among the others.

In Study 4 (601 participants, of whom 120 were excluded for failing the check that there was no reward in the test, as specified in the preregistration), each person was randomly assigned to one of three conditions and completed 80 training trials: on each one, he or she pressed a key to share or not share a headline and saw “won +5 points” or “won 0 points.” In accuracy training, the participant earned 5 points for sharing a true headline or for not sharing a false one; in misinformation training, the reverse; in the control condition, there was no feedback. He or she then decided about 16 new headlines, eight true and eight false, as he or she would on Facebook and without reward. In the test, without feedback, the training pattern persisted: after accuracy training, 66% of the true headlines and 24% of the false ones were shared, versus 40% and 24% in the control condition and 54% and 33% in misinformation training.

Habit, as the authors use the term, is a rewarded response that is repeated in a context, becomes associated with recurring cues, and comes to occur with little sensitivity to outcomes (Ceylan et al., 2023). In operant terms, it is a response maintained by reinforcement and under the stimulus control of recurring features of the context, here the format of the post and the share button, that persists when the reinforcer leaves the scene. What Study 4 adds is that truthfulness can be arranged as a contingency: with points for accuracy, the difference between true and false headlines shared increased and persisted without feedback, and misinformation training raised the overall frequency of sharing without wiping out baseline discrimination. The authors conclude that habit emerged as the main influence on sharing, above political bias and critical reflection.

Monetization and votes are also consequences for the producer, but no verified study measures revenue or votes as an outcome of posting; they remain as hypotheses in Section 9.7.

9.3 What the audience does and what maintains it

The audience also behaves, and it is in the audience that the producer's reinforcement originates. What the verified studies show is, above all, what the audience does in the face of each type of content. Rathje et al. (2021) analyzed 2,730,215 Facebook and Twitter posts. In Study 1, these were accounts of news outlets classified as liberal (for example, the New York Times and MSNBC) or conservative (for example, Fox News and Breitbart), with 599,999 Facebook posts and 227,229 tweets; in Study 2, which replicated the first in another context, they were accounts of members of the U.S. Congress, Democrats and Republicans, with 1,078,562 tweets (up to 3,200 from each member) and 825,424 Facebook posts. The data were collected in 2020.

To determine whether a post was about the poster's own political camp (the in-group) or the opposing one (the out-group), the authors counted, in each text, the words from three lists: the 100 most famous Democratic and Republican politicians according to YouGov, all members of Congress from both parties, and about ten political identity terms (such as "liberal" and "conservative"). A tweet with two words from the Republican list, for example, received a score of 2 in that category. For a conservative account, conservative words counted as in-group language and liberal words as out-group language; for liberal accounts, the reverse applied. Using the same method, they counted words of negative affect, positive affect, and moral-emotional language, with dictionaries already validated in other studies, in order to compare the weight of out-group language with that of these emotions. In each dataset, a regression related the number of shares (log-transformed) to these counts, controlling for having a link, photo, or video, the account's number of followers, and whether the tweet was a retweet; the same regressions were applied to each Facebook reaction. The result of each regression is a coefficient per word, which the authors express as a percentage increase.

Each out-group term was associated with an increase of 35 to 57% in diffusion for news outlet accounts and of 65 to 180% for congressional accounts, depending on the dataset. The average across the eight datasets, in a random-effects meta-analysis conducted by the authors, was about 67% per word (41% for news outlets and 99% for Congress), an effect about 4.8 times that of negative-affect language and 6.7 times that of moral-emotional language. The authors speak of an increase in the odds of the post being shared; the model estimates, more precisely, the ratio between the expected numbers of shares. Posts that referred to the out-group were shared about twice as often as those that referred to the in-group. Facebook offers six reactions ("like," "love,"

"haha," "sad," "wow," and "angry"): "angry" was the most popular in seven of the eight datasets, by the authors' count, and out-group language predicted "angry" reactions, while in-group language predicted "love" reactions. Neither political orientation nor platform moderated the out-group effect, which was stronger among political leaders than among news outlets.

The authors read the result in terms of social identity: messages that satisfy group identity motives may receive more engagement, and posts about the out-group may succeed because they appeal to anger, outrage, and mockery (Rathje et al., 2021). This is a reading from another framework. In operant terms, what the study establishes is that certain properties of content, here reference to the out-group, are associated with more sharing and reacting responses; why these responses are reinforced for those who emit them is what it does not measure.

Guerin (1994) offers a neighboring hypothesis for this part of the cycle. Exchanging attitudes can function as a ritual social event, in which the reinforcement lies in the exchange itself and not in what is said, and the audience is a discriminative stimulus, so that what is declared tends to meet the contingencies of the community that occasioned the declaration. For the audience of a social network, this suggests that liking, sharing, or commenting may be, before it is an evaluation of the content, part of the exchange that maintains the community. This is a hypothesis: the studies cited do not measure the consequence that the audience itself receives.

DATA	The like rate is associated with a shorter interval between posts (Lindström et al., 2021); a sharing pattern established through feedback persists when the feedback is withdrawn (Ceylan et al., 2023); posts about the political out-group are shared about twice as often as those about the in-group (Rathje et al., 2021).
INTERPRETATION	Posting and sharing are operants sensitive to social feedback and to the platform context, and the content to which the audience responds most is, in part, what the producer comes to produce.
HYPOTHESIS	The audience's sharing and reacting are maintained by the feedback it receives from its own network. None of the studies measures this.

9.4 The cycle: each side shapes the other

The cycle is the interlocking of two three-term contingencies, the producer's (Section 9.2) and the audience's (Section 9.3). The producer's contingency has as its occasion the audience, the network's norm, and the platform's interface; as its response, posting a piece of content; and as its consequence, likes, shares, and reach. The audience's contingency has as its occasion the content that appears in the feed; as its response, liking, sharing, commenting, or ignoring; and, as its consequence, what the studies do not measure and Section 9.3 left as a hypothesis. What makes them a cycle is that each contains, as one of its terms, the other's response: the post is the audience's occasion, and the audience's reaction is the producer's consequence. This is what Glenn (2004) calls interlocking behavioral contingencies.

CONCEPT IN USE

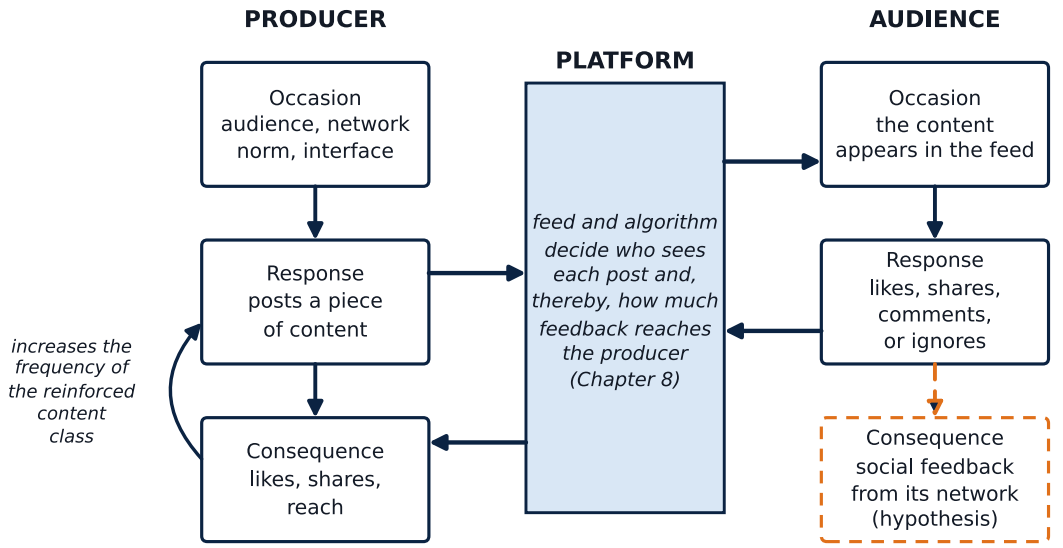
Interlocking behavioral contingencies (Glenn, 2004), revisited from Chapter 5

Chapter 5 (Section 5.1) presented the concept: contingencies interlock when what each of two or more people does serves as an environmental event for what the others do (Glenn, 2004), and the outcome is not the cumulative effect of each person behaving alone but the effect of their interrelated behavior. In Chapter 5 the participants were speaker, audience, and target group; here they are the producer and the audience, and between them stands the platform.

The figure is not a finding; it is a model. Two of its arrows have data in the chapter's studies: the one that runs from feedback to production (Lindström et al., 2021; Ceylan et al., 2023; Brady et al., 2021, Section 9.5) and the one that links the type of content to the audience's reaction (Rathje et al., 2021). The one that closes the cycle, in which what the audience sees changes what it then does, has indirect support: Brady et al. (2021) show, in a simulated environment, that what appears in the feed functions as a norm and that the norm alters expression (Section 9.5). No study measures both halves in the same population and over time. The middle band is the platform, which decides who sees what and, therefore, how much feedback reaches the person who posts; the verified sources say little about it, and Chapter 8 deals with it. Here it is the middle term.

FIGURE 9.1

The producer–audience cycle: two interlocking three-term contingencies



Note. Original diagram by the author, without data. Solid line: relations with data in the chapter's studies; dashed: hypothesis. The platform decides how much feedback reaches the producer (Chapter 8).

9.5 Moral outrage learned through reinforcement: the evidence

Among the studies read, the one that comes closest to the cycle of Section 9.4 is that of Brady et al. (2021), on the expression of moral outrage. The authors start from two hypotheses: that people change what they express through reinforcement learning, adjusting their expressive behavior to the social feedback they receive, and through norm learning, adjusting their expression to what they infer to be normative among their peers, by observation. The design has two parts.

In the first, two preregistered observational studies gathered the complete tweet histories of "politically engaged" users, identified by having tweeted at least once about the confirmation hearings of a Supreme Court justice in October 2018 (Study 1, 3,669 users), or about an incident in which an airline mistreated a passenger (Study 2, the same number); the authors count 7,331 users in all, with 12.7 million tweets. To tell whether a tweet expressed moral outrage, they defined it by three components (feelings in response to a perceived violation of personal moral principles; emotions such as anger, disgust, and contempt; and reactions such as blaming, holding responsible, or wanting to

punish) and trained a classifier on 26,000 tweets annotated on the basis of this definition. The classifier had an accuracy of 75% (the proportion of tweets classified correctly) and an *F*1 of .71 (a score from 0 to 1 that combines how many of the outrage tweets it found and how many of those it flagged as outrage really were).

Next, for each user, they related the amount of outrage on a given day to the feedback (likes and shares) that the previous day's outrage had received, controlling for tweeting frequency, number of followers, the presence of links and media, the previous week's history, and any feedback not tied specifically to outrage (to separate the effect of feedback on outrage from simply tweeting more when one receives more feedback). Daily outrage expression was positively associated with the feedback received by the previous day's outrage (regression coefficients *b* of 0.03 and 0.02 in Studies 1 and 2). In the authors' example, a user who averaged five likes or shares per tweet and received ten when expressing outrage (a 100% increase) would tend to increase the next day's outrage by 2 to 3%, an effect the authors describe as small and capable of scaling up. The longer the user's history, the smaller the effect of recent feedback, and users embedded in more ideologically extreme networks expressed more outrage and responded less to feedback.

In the second part, two preregistered experiments (120 participants each, recruited on Prolific, all liberals) simulated Twitter. In Study 3, participants were randomly assigned to one of two norm conditions and began by scrolling through a feed of 12 tweets from a "new network," whose preferences they were supposed to learn. When the norm was outrage, 75% of the feed's tweets expressed outrage and displayed more likes and shares than the neutral ones; in the neutral-norm condition, all were neutral. They then completed 30 trials in which they chose, between two political tweets, one outrage and one neutral, which one to "post" to the network, and received likes that, they were told, came from previous participants; the goal was to maximize these likes. Choosing outrage earned more likes, on average, in both conditions. Study 4 used the same paradigm with one difference: the greater feedback went to the tweet congruent with the norm, that is, to outrage in the outrage-norm condition and to the neutral tweet in the neutral-norm condition.

The authors measured two processes. Norm learning is the tendency to choose, on the first trial, the type of tweet that the feed showed to be normal. Reinforcement learning is the tendency to choose increasingly, across trials, the type of tweet that earned more likes. The results are reported as odds ratios (*OR*): an OR of 1 means equal odds, and above 1, an advantage for the choice in question. On the first trial, those who saw the outrage norm chose the outrage tweet with much greater odds (OR = 4.94 in Study 3 and 5.38 in

Study 4), and those who saw the neutral norm chose the neutral one (OR = 1.73 and 1.54). Across trials, choice of the reinforced tweet increased (in Study 3, OR = 1.10 across both conditions; in Study 4, 1.03 for outrage and 1.06 for neutral tweets). And the effect of reinforcement was smaller among those who saw the outrage norm (OR of 0.95 in Study 3 and 0.97 in Study 4 for the difference between conditions; below 1, the effect is smaller in that condition), which the authors read as less dependence on feedback for those who already saw outrage as the norm.

The authors' terms come from computational reinforcement learning (feedback, prediction error). What matters here is the functional relation: the expression of outrage varied with the social consequences that preceded it, in natural data and in an environment in which feedback was arranged. What reinforcement alone does not explain, the authors say, is that the feed shows what is normative in the network and that those who observe it depend less on feedback. In behavior-analytic language, this comes close to control by modeling and by rules (a reading proposed here), and it is not the same as direct reinforcement. The link to algorithms is the authors' interpretation: because they determine how many users see a post, algorithms affect feedback and may influence moral behaviors by exploiting reinforcement learning tendencies; and design decisions aimed at profit through engagement may affect it indirectly, because content that provokes outrage attracts high engagement. None of the four studies measures design decisions.

The authors themselves record the limits: the observational studies do not allow causal inference, the users were identified by having tweeted during episodes of public outrage, and the conclusions concern expression in text, not emotion (Section 9.8).

DATA	In 12.7 million tweets from 7,331 users, the next day's outrage increased by 2 to 3% for each 100% increase in the feedback to the previous day's outrage; in two experiments with 120 participants each, the choice of outrage tweets increased with feedback and followed the displayed norm (Brady et al., 2021). In more than one million posts, the interval between posts decreased with the like rate (Lindström et al., 2021).
INTERPRETATION	The expression of outrage, like posting frequency, is sensitive to social feedback; the network context, observed in the feed, modulates this sensitivity.

HYPOTHESIS	On platforms where the algorithm amplifies what provokes engagement, feedback for outrage grows and expression grows with it. This is the authors' hypothesis, taken up here; no study has manipulated the algorithm and measured outrage.

9.6 The politician as content producer

A politician who posts on social media is a producer with a particular audience: followers with their own reinforcement history on the topic, potential voters, the press. Rathje et al. (2021) break down the analysis by type of account and by political camp, which makes it possible to read the cycle for this producer. In 1,078,562 tweets and 825,424 Facebook posts from Democratic and Republican members of Congress, those that referred to the out-group were shared more in both camps. The estimated increase per out-group term, in the authors' meta-analysis, was 99% in the congressional datasets, versus 41% in the news outlet datasets, and the effect was not moderated by political orientation. The study covers the years leading up to the 2020 election, is correlational, and the authors say they do not know how algorithmic choices contribute.

The finding is about accounts and content: it says what received more sharing, not why each politician posted what he or she posted. The study applies the same criterion to both camps, which is what the method followed here requires. Applying it to Brazilian politicians is the question one would like to ask; none of the verified sources answers it.

DATA	Posts by members of the U.S. Congress that referred to the political out-group were shared more than those that referred to the in-group, on Facebook and Twitter, in both camps, with an effect larger than that observed for news outlets (Rathje et al., 2021).
INTERPRETATION	The feedback the audience gives a political producer favors, in the data, content about the out-group, which describes the producer's contingencies and does not pass judgment on people or parties.
HYPOTHESIS	The same holds for Brazilian politicians and for other platforms. No verified source tests this.

9.7 The metacontingency of the ecosystem: producer, audience, platform, and advertiser

So far the chapter has stayed at the level of the behavior of individuals in interlocking contingencies. The question that remains is whether the whole, made up of producers, audiences, platforms, and those who pay for attention, sustains itself, as a practice that recurs because its effects select it. Behavior analysis has a concept for this, the metacontingency. The chapter uses it as a hypothesis and with the precautions the literature itself calls for.

CONCEPT IN USE

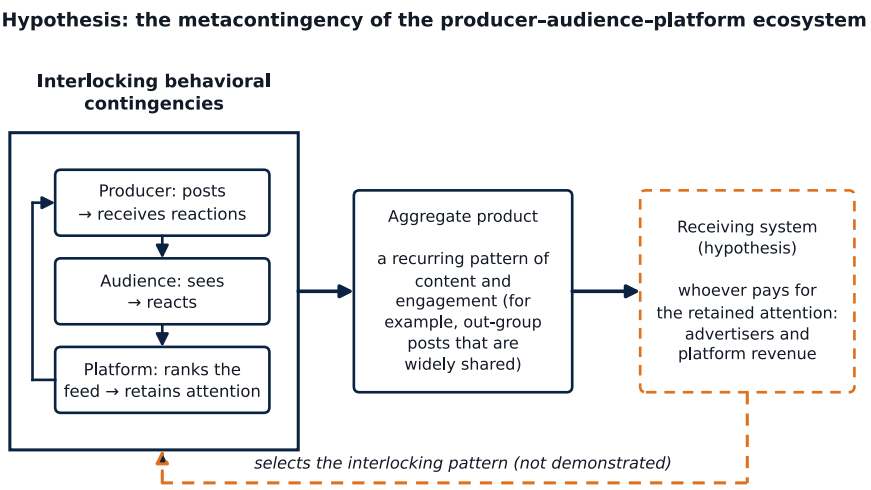
Metacontingency (Glenn, 1988, 2004)

The metacontingency is the unit of analysis that takes in a cultural practice, with all its variants, together with the aggregate result that those current variants produce (Glenn, 1988). What cultural outcomes select is not individual behavior but the interlocking behavioral contingencies of which the practice is made. In Glenn (2004), the lineages that evolve are these recurring contingencies, which function as an integrated unit and have an outcome that affects the probability of their future recurrence. The formula "interlocking behavioral contingencies, their aggregate product, and their receiving system" is from Glenn and Malott (2004, as cited in Houmanfar & Rodrigues, 2006). A macrocontingency is something else: the relation between a practice and the aggregate sum of the consequences of the macrobehavior that constitutes it, cumulative effects that cannot control the behavior of individuals as operant consequences do (Glenn, 2004) and that do not involve selection at the cultural level.

Applied to the ecosystem, the scheme requires three things: recurring interlocking contingencies, an aggregate product, and a receiving system that, by receiving the product, maintains the recurrence of what generated it. Figure 9.2 proposes how to fill these in. The interlocking contingencies are those of Section 9.4, with the platform as a third term: the producer posts and receives reactions, the audience sees and reacts, the platform ranks the feed and retains attention. The aggregate product would be a recurring pattern of content and engagement, for example out-group posts with a great deal of sharing. The receiving system would be whoever pays for the retained attention: advertisers and, through them, the platform's revenue.

FIGURE 9.2

Hypothesis: the metacontingency of the producer–audience–platform ecosystem



Note. Original diagram by the author, without data: it is a hypothesis, not a result. The three terms follow the formula of Glenn and Malott (2004, as cited in Houmanfar & Rodrigues, 2006). Houmanfar and Rodrigues propose another configuration: cultural milieu, aggregate product, receiving system.

The verified sources support the interlocking contingencies in part (Sections 9.2 to 9.5) and leave the second half of the hypothesis as an assumption of their authors: Brady et al. (2021) state that content that provokes outrage attracts high engagement, which may affect moral behavior when design aims at profit through engagement; Rathje et al. (2021), that the business model of social media is based on generating engagement, which leads to advertising revenue; Ceylan et al. (2023), that habitual users are central to profit models built on advertising. These are interpretive premises, not results: none of these studies measured revenue or design decisions.

Two precautions come from the literature itself. The first is the distinction between metacontingency and macrocontingency. If what we want to explain is a cumulative effect, such as polarization (an example chosen here), what we have is a macrocontingency: the aggregate effect of many independent operants cannot function as a consequence that controls each individual's behavior (Glenn, 2004). What keeps the user posting is still the feedback he or she personally receives. Cultural engineering, Glenn (2004) writes, requires pinpointing the macrocontingencies responsible for undesirable effects and then changing whichever operant contingencies or metacontingencies are relevant; Part IV returns to this. The second precaution is that the concept is debated. Houmanfar and Rodrigues (2006) propose the configuration cultural

milieu, aggregate product, and receiving system, with the aggregate product as what is selected, because they judge that the current form mixes levels of analysis, and they say their ideas are preliminary. For the ecosystem the difference matters: if what the receiving system selects is the aggregate product, and not the interlocking contingencies, the empirical question is which products recur (what kind of content and engagement), not which arrangements of producer, audience, and platform.

DATA	The verified studies show sensitivity to social feedback in posting, sharing, and expressing outrage (Sections 9.2 to 9.5) and more sharing of content about the out-group (9.3 and 9.6). Their authors state, without measuring it, that the platforms' business models depend on engagement.
INTERPRETATION	The producer's and the audience's contingencies are interlocking in Glenn's (2004) sense, with the platform as the mediator of feedback.
HYPOTHESIS	The interlocking pattern recurs because its aggregate product is received by advertisers who pay for attention. This would require showing that variation in the aggregate product is followed by variation in revenue and that revenue maintains the practice. No verified source does so.

9.8 Rival interpretations and limits

The cycle has rival readings, and what separates one from another depends on the type of data. The platform data are of two kinds, with different limits. The observational data (Lindström et al., 2021; Brady et al., 2021, Studies 1 and 2; Rathje et al., 2021) describe associations in self-selected populations; the experimental data (Lindström et al., 2021, Study 3; Brady et al., 2021, Studies 3 and 4; Ceylan et al., 2023) manipulate feedback in simulated environments, with participants from research platforms, and measure choices, not posts on real networks.

Four limits apply to the entire chapter. First, no source is Brazilian, and the platform sources are from the United States or from online research samples; Brady et al. (2021) say they do not know whether the findings generalize to other populations, to other platforms, or outside the U.S. political context, and that Twitter users are not representative. Second, the measured effects are small: a 2 to 3% increase in the next day's outrage for each 100% increase in feedback (Brady et al., 2021) and a difference of about 10.9% in latency (Lindström et al., 2021). Small effects can scale up, as the authors note, but that is

an argument, not a result. Third, the experiments from 1956 to 1969 use students and, in Scott (1957), an interval of five minutes. Fourth, the cycle as a whole is a model: each arrow has partial support, and the closing of the cycle over time, with real producers and audiences, has not been measured.

Rival interpretations for the same set of data

Exposure determined by the algorithm. What a post receives depends on how many users see it, and that is decided by the platform (Brady et al., 2021). An association between content and reactions may reflect what the algorithm showed, not what the audience would choose; Rathje et al. (2021) say they do not know how algorithmic choices contribute.

Selection of the producer and the network. Users who express more outrage may follow more extreme accounts, rather than learn to express it in the network they are in; the authors acknowledge that network composition may be endogenous (Brady et al., 2021, Studies 1 and 2).

Other motives for posting. There are many motives besides reward, such as self-expression and relationship building (Lindström et al., 2021).

Norm and identity instead of reinforcement. What is observed may be conformity to network norms (Brady et al., 2021) or the expression of group identity (Rathje et al., 2021), without individual feedback being what maintains the response.

Expression without emotion. The conclusions concern expression in text, not emotion, which was not measured; the authors acknowledge that platform design may decouple the one from the other (Brady et al., 2021).

From the telephone to the network. The experiments from 1956 to 1969 use a known listener, a verbal reaction, and few interactions; social networks have multiple audiences, counted and visible feedback, and long cycles. The step from one to the other is made by analogy, and here it is acknowledged as such.

For those who want to use the chapter in the analysis of statements (Chapters 3, 5, and 7), the consequence is practical. The listener's reaction is a plausible and measurable controlling variable of what the speaker says, but claiming that a specific politician's statement was maintained by a particular audience requires a record of that audience and its reactions, which a fact-check, by itself, does not provide.

Summary

What is said is sensitive to what happens after it is said, and this was shown with declared opinions before social networks existed: with “Good” after answers, with a winning vote after a debate. The effect is small, appears outside the laboratory, depends on the reinforcer and on the task, and awareness of

the contingency is disputed (Section 9.1). On platforms, posting frequency tracks the like rate, sharing becomes only weakly sensitive to truthfulness when it is established through repetition, and the expression of outrage tracks feedback and the norm displayed in the feed, with small effects in the observational data (Sections 9.2 and 9.5). On the audience's side, we know what it does in the face of each type of content; the consequence that maintains it is a hypothesis (9.3). The cycle joins the two halves as interlocking contingencies, has partial support for each arrow, and no study measures it as a whole (9.4).

For the politician as producer, the only verified data are those on members of the U.S. Congress: content about the out-group is shared more, in both camps, and more among members of Congress than among news outlets (9.6). Extending this to Brazil is a hypothesis. At the cultural level, the ecosystem can be read as a metacontingency, with platform and advertisers as terms, but this requires an outcome that no verified source measures, and part of what is attributed to it, such as polarization, is a cumulative effect, a macrocontingency (9.7).

Chapters 3, 5, and 7 use from this chapter the idea that the audience's reaction is a controlling variable of speech; Chapter 8, the platform band; Chapter 12, the distinction between acting on the operant contingencies of the producer and the audience and acting on the metacontingencies.

References

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), Article eabe5641. https://doi.org/10.1126/sciadv.abe5641

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Glenn, S. S. (1988). Contingencies and metacontingencies: Toward a synthesis of behavior analysis and cultural materialism. *The Behavior Analyst*, *11*(2), 161–179. https://doi.org/10.1007/bf03392470

Glenn, S. S. (2004). Individual behavior, culture, and social change. *The Behavior Analyst*, *27*(2), 133–151. https://doi.org/10.1007/bf03393175

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Hildum, D. C., & Brown, R. W. (1956). Verbal reinforcement and interviewer bias. *The Journal of Abnormal and Social Psychology*, *53*(1), 108–111. https://doi.org/10.1037/h0040518

Houmanfar, R., & Rodrigues, N. J. (2006). The metacontingency and the behavioral contingency: Points of contact and departure. *Behavior and Social Issues*, *15*(1), 13–30. https://doi.org/10.5210/bsi.v15i1.342

Insko, C. A. (1965). Verbal reinforcement of attitude. *Journal of Personality and Social Psychology*, *2*(4), 621–623. https://doi.org/10.1037/h0022485

Insko, C. A., & Melson, W. H. (1969). Verbal reinforcement of attitude in laboratory and nonlaboratory contexts. *Journal of Personality*, *37*(1), 25–40. https://doi.org/10.1111/j.1467-6494.1969.tb01729.x

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), Article 1311. https://doi.org/10.1038/s41467-020-19607-x

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Scott, W. A. (1957). Attitude change through reward of verbal behavior. *The Journal of Abnormal and Social Psychology*, *55*(1), 72–75. https://doi.org/10.1037/h0040249

Spielberger, C. D., Levin, S. M., & Shepard, M. G. (1962). The effects of awareness and attitude toward the reinforcement on the operant conditioning of verbal behavior. *Journal of Personality*, *30*(1), 106–121. https://doi.org/10.1111/j.1467-6494.1962.tb02308.x

instituto WALDEN4 | ACADEMIA WALDEN4

TV WALDEN4

Science on video, on the air all day long

Classes, lectures, and conversations about psychology, autism, and behavior analysis, on themed channels that play nonstop, synchronized for everyone.

667 videos on 10 channels

Free access, no sign-up required

walden4.com.br/tv

CHAPTER

10

The Overton Window: What Can Be Said Without Being Punished

An experiment by Bursztyn et al. (2020), conducted in 2018, serves as an entry point to the Overton window. The 1,600 residents of the Pittsburgh metropolitan area who took part all received the same offer: a cash bonus if they authorized a donation, in their name, to a strongly anti-immigration organization. At random, some learned that Trump had won the metropolitan area in the 2016 election; others, that Clinton had carried the county in which Pittsburgh lies. Half of each group was led to believe that the decision would be anonymous; the other half, that it would be visible to other residents of the area. In the anonymous condition, 31% (Trump won) and 30% (Clinton won) accepted. In the visible condition, 33% and 20% did (Bursztyn et al., 2020).

Only one piece of information varied, and it said nothing about immigration: it said who had won the election in the neighborhood. What varied with it was what people agreed to do when others might see. It is this shift in what can be done or said in public at no cost that political conversation calls a shift in the Overton window.

The chapter asks what the concept means in terms of contingencies and what the verified data allow us to say about who shifts the window and why. Section 10.1 presents the concept, its origin, and the translation proposed here; 10.2 places it in the relation between speaker and audience; 10.3 returns to the Pittsburgh experiment; 10.4 deals with two elections, Brazil's in 2018 and the U.S. election of 2016; 10.5, with politicians and social media; 10.6, with what holds the edge in place; 10.7 proposes questions for situating a statement; and 10.8 gathers the rival interpretations.

Three cautions in reading the chapter. The first is that the concept was born in a think tank that uses it to promote its own policies and is, in its origin, a metaphor about public policy, not a definition of verbal behavior. The second is that the data measure different things (perception, public action, classified text), and no verified study links a politician's statements to those of his or her voters with a design that identifies the direction of influence. The third is that the election studies deal with the victories of Trump, in 2016, and of Bolsonaro, in 2018; none deals with a mirror case, the victory of a candidate from the opposite camp. The chapter's criterion applies to any camp, but what has been shown does not allow us to claim symmetry. Throughout, data, interpretation, and hypothesis are kept separate.

10.1 The concept, its origin, and the proposed translation

The original formulation is that of the Mackinac Center for Public Policy, which maintains a page on the concept. According to the page, the Overton window is a model of how a society's ideas change over time and of how that change bears on politics. Its central idea is that politicians are constrained in what they can back: as a rule they adopt only policies that are widely accepted as legitimate, and these are inside the window; other policies exist, but whoever defends them stands to lose popular support, and they are outside it (Mackinac Center, 2019).

The window can shift and widen. On occasion politicians move it themselves by backing a policy that sits outside it, though seldom; usually the window moves as society's values and norms slowly evolve, a process that cannot readily be steered from the top. The page concludes that politicians are more followers than leaders, and that the institutions that set and reinforce norms (families, workplaces, friends, media, churches, associations, schools) do more to shape politics than they are usually given credit for. The example is Prohibition: federal policy a few generations ago, it is now something no politician proposes bringing back, and it has left the window. Of interest here is a passage from the frequently asked questions: the window reflects what society believes, and this can be influenced as much by facts as by inaccurate or misleading information; the page answers that falsehoods, distortions, or misunderstandings can indeed shift the window, adds that spreading misleading information on purpose is plainly wrong, and gives as examples of errors that shifted the window the underestimate of Medicare costs and the belief in Iraq's weapons of mass destruction (Mackinac Center, 2019).

The reader should read this knowing where it comes from. The concept was developed in the mid-1990s by Joseph P. Overton, who was senior vice president of the center when he died in 2003, and the center says it employs the window to help advance its own policy agenda. The page admits, finally, that the model does not account for every aspect of how politics works: it describes one thing, namely that politicians back only the policies they judge will not damage their chances at the polls (Mackinac Center, 2019). The description is not invalidated by this, but it comes from an interested party and not from a study, and the chapter uses it as a starting point, not as evidence.

As a formulation for a science of behavior, the concept has one gap and one substitution to make. The gap is operational: the page does not say how to measure "widely accepted" or "popular support," so that, as formulated, what is inside or outside the window is known only afterward, from what politicians did. The substitution is one of unit and criterion. The unit is no longer the public policy that a politician can support but the statement that a speaker can emit. The criterion is no longer "popular support" but the consequence the statement receives from the audiences that matter to the speaker, for which the analysis of verbal behavior already has a concept: the audience that punishes. The box gathers what Skinner (1957) says.

CONCEPT IN USE

Negative audience and nonpunishing audience (Skinner, 1957)

Skinner (1957) separates the listener who simply fails to reinforce verbal behavior from the listener who punishes it: an audience in whose presence verbal behavior is punished can be called a "negative audience," and kings, senior government officials, and powerful executives can be effective negative audiences. Some negative audiences are tied to topics (one learns to avoid mentioning certain subjects or events), and some control only part of the speaker's repertoire, such as the community that punishes the language spoken in the child's home. A negative audience's effect shows up only when a positive audience is also present, because it takes the form of a weakening of the behavior appropriate to the positive one; when the two are together, a dilemma arises, in which the response is either emitted and punished by one or withheld, with a loss of prestige before the other, and as long as the audiences remain separate the dilemma does not arise. The seditious orator who sees a policeman approaching speaks less and less as the negative audience gains relative weight, perhaps even dropping below the overt level. The opposite is the nonpunishing audience: the psychotherapist must establish himself as such, and if the control exerted by the audience changes, the patient may emit previously punished behavior; the confidant is the sympathetic person to whom one speaks with less fear of punishment than to random listeners.

From this we take the definition used in the rest of the chapter. It is proposed here, and is neither Skinner's nor the Mackinac Center's.

PROPOSED DEFINITION

An audience's window.

For a given speaker and topic, the set of statements that this speaker emits in the presence of this audience without the audience punishing them. "Punish" has Skinner's functional sense: it is the consequence that makes the statement less probable, which includes loss of prestige, and not only formal sanction.

Two consequences follow from the definition. The first is that the window is a property of the relation among speaker, audience, and topic, and not of the statement: the same statement is inside one audience's window and outside another's, and there is no single window of a society; there are as many as there are relevant audiences. The window that political conversation talks about, that of "public debate," is, in the reading proposed here, the window of the audiences that dominate an arena at a given moment, and the analysis needs to say which ones. The second is that "inside" and "outside" are not categories of what is true, fair, or right: they say what is or is not usually punished by those who listen. A false statement can be inside one audience's window, and a true one outside another's. The definition takes no side on either.

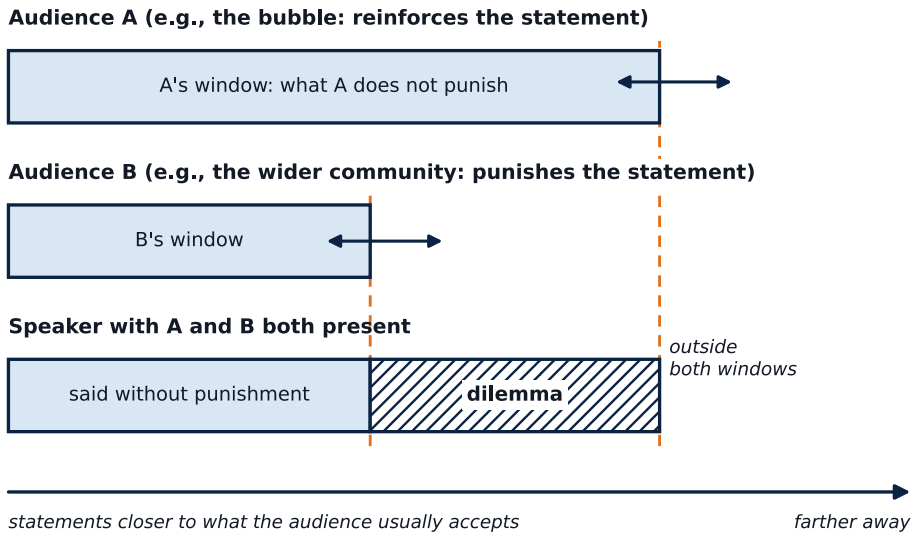
DATA	The Mackinac Center describes the window as the set of policies that politicians can support without losing popular support, says that it moves mainly through the slow evolution of values and norms, and that politicians are more followers than leaders; the center itself says it uses the concept to promote its policies (Mackinac Center, 2019). Skinner (1957) describes the negative audience, the dilemma of audiences with different windows, and the nonpunishing audience.
INTERPRETATION	An audience's window is the set of statements that it does not punish, for a given speaker and topic; it is a relation, not a property of the statement.
HYPOTHESIS	Politicians "follow" to the extent that what they emit is controlled by punishment from the audiences that matter to them. Testing this requires measuring the sanction, not just what was said; the Mackinac Center page does not say how.

10.2 From the policy window to the statement window: what shifts the edge

Defining the window as a property of the relation among speaker, audience, and topic (Section 10.1) has an implication that Figure 10.1 depicts. Two audiences have windows of different widths for the same topic: A's, for example a bubble (Chapter 6) in which the statement is reinforced, is wide, and B's, for example a broader community, is narrow. When only A is present, the speaker emits up to A's edge; when only B is present, up to B's. When both are present, the band between the edges is Skinner's dilemma zone: the statement is reinforced by A and punished by B, and the speaker either emits it and is punished by B or withholds it and loses prestige with A. What he or she does depends on the relative weight of each audience, which the figure does not represent.

FIGURE 10.1

The window is relative to the audience: two windows and the dilemma zone



Original conceptual diagram, no data. Double arrows: the edge shifts (Section 10.2). Hatched: reinforced by A and punished by B (Skinner, 1957, pp. 230–231). Simplification: both windows assume the same ordering of statements.

Note. Original diagram, without data. The double arrows indicate that the edge shifts (see the text). The hatched band is the dilemma zone (Skinner, 1957). Simplification: both windows assume the same ordering of statements, from most to least accepted; in reality, two audiences may order statements differently.

In terms of contingencies, the edge of a window shifts when the probability of punishment for the statements adjacent to it changes. Seven candidates to explain the change are gathered here, each with what the verified sources provide. The first is who makes up the audience: the entry or exit of listeners

with a different history changes what is reinforced and what is punished, and Chapter 6 dealt with bubbles as audiences that reinforce what the group accepts. The second is the audience's power of sanction: Skinner (1957) speaks of "effective" negative audiences, and countercontrol (Delprato, 2002), discussed in Section 6.4, is the set of responses that reduces the effectiveness of aversive control; without power of sanction, the edge recedes. The third is information about what others tolerate, the subject of Sections 10.3 and 10.4. The fourth is what the environment shows that others do, that is, descriptive norms, discussed in Section 10.6. The fifth is what receives feedback at the edge: borderline statements that are reinforced come to be emitted more often, and Section 10.5 deals with this with data on politicians. The sixth is repetition: repeatedly encountering a false headline reduced how unethical it seemed to publish and share it, in four experiments (Effron & Raj, 2020), a result that Chapter 3 (Section 3.5) discusses at greater length. The seventh is reach: the platform decides how many people hear each emission (Chapter 8), and a wide window of a small audience and a narrow window of an enormous audience have very different effects on what circulates.

DATA	Each of the variables above has support in at least one verified study, of very different kinds (experiments, tweet series, an experiment with repeated headlines), and none has been manipulated together with the others.
INTERPRETATION	The same change in speech can be produced by more than one of them, and the term "shift of the window" describes the result, not the mechanism.
HYPOTHESIS	The seven variables act jointly in practice (for example, information about others' tolerance arrives through a platform, which decides reach). No verified source separates their contributions.

10.3 Information about what others tolerate: the Pittsburgh experiment

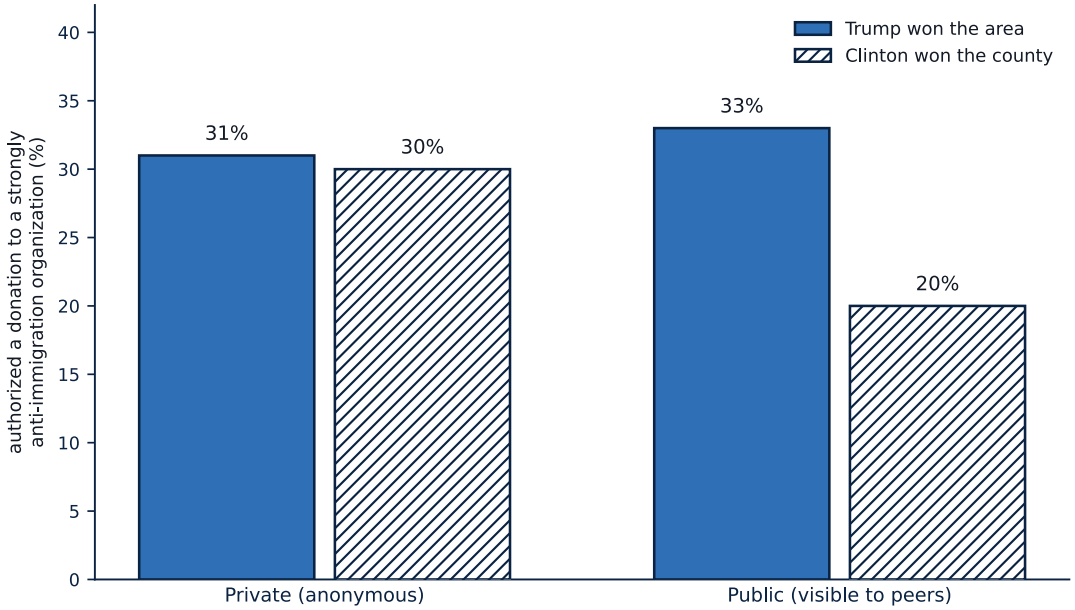
The experiment from the opening is the most direct test, among the verified sources, of the hypothesis that information about what others accept shifts what people do in public. Bursztyn et al. (2020) carried it out in two experiments that look at both sides of the interaction: those who act and those who observe and can sanction. In the first, from August to October 2018, an online panel company recruited 1,600 participants from the seven counties of the Pittsburgh metropolitan area. At random, half were informed that Trump had won the metropolitan area in the 2016 election, and the other half that Clinton

had won Allegheny County, where Pittsburgh is located. A cash bonus was then offered to those who authorized a donation, in their name, to a strongly anti-immigration organization (for the authors, accepting was a profitable xenophobic action), and half of the participants were assured of anonymity (private condition), while the other half were led to believe that the decision would be visible to peers in the area (public condition).

In private, donation rates were 31% (Trump won) and 30% (Clinton won), with no significant difference ($p = .894$); in public, 33% and 20% ($p < .001$). The difference in differences, that is, how much the effect of the information in public exceeds its effect in private, was significant ($p = .008$). The mean perceived local popularity of anti-immigration views was 50.1% among those who learned that Trump had won and 42.6% among those who learned that Clinton had won. In the instrumental variable estimation, which uses the random assignment of the information to estimate the effect of perception on donation, one additional percentage point in this perception increased public donation by 2.3 percentage points and did not change private donation, a result the authors themselves treat as suggestive, because the treatment may act through other channels, although they consider these unlikely.

FIGURE 10.2

Donation to an anti-immigration organization, in private and in public, by who won the election in the area



Note. Values rounded, as in the article (Bursztyn et al., 2020). N = 1,600, Pittsburgh metropolitan area, 2018; the information about the winner was randomly assigned. The treatment is information about who won the election in the area, not a candidate's statement.

Figure 10.2 shows where the difference lies, and it is subtler than "Trump's victory unleashed xenophobia." The authors observe that the result comes mainly from the wedge between donating in private and in public in the Clinton condition (30% versus 20%), with no significant difference in the Trump condition (31% and 33%; $p = .594$). They read this as an indication that, in Pittsburgh, Trump's victory did not generate significant pressure for tolerant people to act xenophobic in public, whereas learning of Clinton's victory was associated with social pressure for xenophobes to appear tolerant in public. Their reading is that the information does not make participants more xenophobic: it makes those who already were more comfortable expressing their views in public.

The second experiment looks at the side of those who observe and can sanction. In November 2018, about 1,800 respondents who identified as Democrats split $2 with another person in a dictator game, in which the one who decides how much to pass on keeps the rest. The other person was a participant in the first experiment who had made the donation to the anti-immigration organization (only those who donated were included), and the respondent did not know that the area was Pittsburgh. Two pieces of information about this donor were crossed at random: who had won the election in the donor's area, Trump or Clinton, and whether the donation had been private or visible to peers, which yields four conditions. The mean amount given was $0.78, $0.81, and $0.78 in the three conditions in which there was no reason to suppose pressure to act xenophobic (the two private ones and the public one where Clinton won) and $0.88 in the condition of public donation in an area where Trump won; the share who gave nothing was 18%, 21%, and 22% in the three and 8% in the last. The authors explain the result by the fact that a public donor in an area where Trump won may be under social pressure to act xenophobic, which weakens the inference that he or she is really xenophobic; that donor therefore receives more, that is, is punished less.

In the language used here, Experiment 1 moves the edge on the speaker's side and Experiment 2 on the side of the listener's sanction, and both edges move with information of the same kind, about who won the election in the area. The information did not change what participants agreed to do where no one would see, and it changed, or ceased to narrow, what they agreed to do where others would see. This is consistent with the definition in Section 10.1, in which the window is a property of the relation with an audience, not of the speaker. The authors adopt a narrow view of social norms, in which they guide public or potentially public actions and not private ones, and they say that what they manipulated was Trump's perceived popularity, not his statements.

The limits are clear. The manipulated variable is information about who won the election in the area, not what a politician said; the outcome is authorizing a donation, not uttering a statement; the sample comes from an online panel in one region and, in Experiment 2, only from Democrats; and there is only one case, on one side of the issue. What the experiment supports is that information about local support for a position alters what people agree to do in public regarding it, without altering what they do in private. That the same holds for statements, for other positions, and for other audiences is an extension proposed here.

DATA	Those who learned that Trump had won the area agreed to donate in public to an anti-immigration organization in a proportion similar to those who would donate in private (33% and 31%); those who learned that Clinton had won agreed less in public than in private (20% and 30%). Democratic observers gave less to public donors in an area where Clinton won than in an area where Trump won (Bursztyn et al., 2020).
INTERPRETATION	Information about the local popularity of a position alters the perceived cost of expressing it in public, for those who act and for those who sanction, without altering the private position. In the terms used here: what shifts is the audience's window, not the speaker's disposition.
HYPOTHESIS	The same holds for statements, for other positions, and for other audiences. The authors do not manipulate statements.

10.4 Two elections, two measures: Brazil, 2018, and the United States, 2016

If information about what others tolerate shifts what people do in public, a surprising election is an obvious candidate for a natural experiment: it informs everyone, all at once, that support for a candidate and for what he says is greater or smaller than imagined. Two verified studies try to read this in social media text and arrive at results that cannot be reduced to each other.

Marino-Fages and Martínez-Pozo (2026) study the 2018 Brazilian presidential election. They treat the result as a surprise: Bolsonaro received 46% of the vote in the first round, well above the polls, which had predicted about 35%, and the surprise varied across municipalities, because support for him varied greatly from place to place. They collected more than 37.6 million tweets in Portuguese, geolocated in Brazil, excluding retweets, posted on Mondays from July 2017 to December 2019, and assigned each tweet to a municipality by lat-

itude and longitude. To measure hate speech they used two methods. The first was a Portuguese language model, BERTimbau (a version of BERT), fine-tuned on 5,668 tweets from 2017 that other researchers had labeled as containing or not containing hate speech. The second was a dictionary of terms that classifies the target of hate into six categories (political hate, political target, homophobia, racism, sexism, and insults). The outcome in each municipality is the share of tweets classified as hate.

The design is difference-in-differences: it compares the change in the share of hate between before and after the election, excluding the period from August to October 2018, and examines how this change varies with Bolsonaro's first-round vote share in each municipality. Identification assumes that, without the election, trends would have been parallel across municipalities with different levels of support for him. For the country as a whole, the average share of hate tweets according to BERT rose from 8% (July 2017 to July 2018) to 9% (2019), about 3,000 more tweets on an average day.

The central result is that hate speech increased after the election and increased more where Bolsonaro received fewer votes: in municipalities with a vote share one standard deviation higher, the increase was 0.033 standard deviations smaller with the BERT classifier and 0.018 with the dictionary (the standard deviation, a measure of each variable's typical variation, puts both on the same scale). The increase was concentrated in homophobia and sexism, themes of the candidate's rhetoric; for general insults, political hate, and political targets, there was no comparable pattern, and the authors conclude that the result is not driven by growing polarization. The number of users who began posting hateful content without having done so before (the extensive margin) grew more where he was less popular, and the intensity of expression among those who had already started (the intensive margin) grew more where he had more support. The share of users posting hate speech rose 10.1% among men, 6.1% among women, and 3.3% among pseudonymous or unidentifiable accounts. As a numerical example, the authors take Rio de Janeiro (58.3% of the vote for Bolsonaro; about 32,000 tweets per weekday, 9% containing hate speech): a similar municipality with a vote share 10 percentage points lower would generate 40 to 50 more hate tweets per day, which they call modest at the individual level but cumulative and broadly participatory.

The authors' interpretation is that the pattern is consistent with a breakdown of the spiral of silence, in which the perception of being in the minority on a values issue leads to self-silencing for fear of social isolation: the election revealed that views associated with the candidate's rhetoric were more widespread than believed, reducing the perceived social cost of expressing them, especially where he was less popular. In the vocabulary used here, this is the

third candidate of Section 10.2, and it coincides with the reading of Bursztyn et al. (2020). In favor of the design, the authors note that Brazil had, during the period, no comprehensive legislative framework on content moderation on platforms, which makes it unlikely that new legal regimes explain the increase. Four caveats apply. The study is a working paper, and its own cover says it has not been peer-reviewed. The classifier had accuracy, precision, recall, and F1 of 77 to 79% on the validation and test sets (accuracy is the proportion of tweets classified correctly; precision, the proportion of hits among those flagged as hate; recall, the proportion of hate tweets it found; F1 combines the last two). Monday tweets and geolocated tweets are treated as representative samples of the universe of tweets, an assumption of the authors, and the ratio of geolocated tweets to the total varies from 4% to 8% by word. And the design measures text, not perceived social cost: the mechanism is the authors' reading, not a result.

Siegel et al. (2021) study the 2016 U.S. election. They assembled two sets of tweets, from June 17, 2015, the day after Trump announced his candidacy, to June 15, 2017: all tweets mentioning Trump or Clinton (more than 600 million about him and more than 150 million about her), collected in real time, so that tweets later deleted or from accounts later suspended remain in the counts; and almost 400 million tweets from a random sample of 500,000 American users, to see whether hate also increased outside the political conversation.

They defined hate speech as hostile and malicious language, motivated by bias, directed at a person or group because of innate characteristics, real or perceived; neither praise for the Ku Klux Klan nor excessive pride in one's own group counts in itself, and white nationalist rhetoric was measured separately. They used two methods. The first was a dictionary of 3,268 terms (slurs against Asians, Black people, immigrants, Muslims, Jews, gay people, and women, plus white nationalist rhetoric), drawn from public databases. Because many terms appeared in usernames, inside other words, or in tweets condemning hate, about 25,000 tweets containing these terms were hand-coded, by three raters each, and used to train classifiers (Naive Bayes) that eliminate these false positives; the hate speech classifier had 94% accuracy, 95% precision, and 90% recall. The second method does not require knowing the words in advance: it measures how much each day's tweets resemble comments from Reddit communities known for racist and alt-right content, taken as hate speech “in the wild,” and only relative comparisons, such as tweets after the election versus those at the start of the campaign, are interpretable.

The analysis is an interrupted time series, which compares the daily level before and after an event, and the authors computed four measures (the number and proportion of tweets and of users with hate per day), which tell essen-

tially the same story. Hateful tweets made up between 0.1% and 0.3% of the total in every month. The authors found no lasting rise in either hate speech or white nationalist language, whether during the campaign or in the wake of Trump's election; key events generated brief spikes, which dissipated quickly, and the largest monthly spike in the Trump data, in late January 2017, was largely misogynistic hate following the travel ban executive order. They say the dictionary is conservative, because it requires the tweet to contain a slur or derogatory term. They limit the scope to Twitter, note that hate speech may have proliferated on specialized platforms, leave the effect of account bans to future research, and warn that seeing a certain kind of speech at a given moment does not mean it has increased, given the bursty nature of Twitter, in which topics gain prominence briefly and return to equilibrium.

The two studies do not contradict each other in any simple way, nor do they add up. They differ in measure (a classifier trained on labeled tweets and a dictionary by target, versus a dictionary that requires a slur and a classification by communities), in design (difference-in-differences across municipalities, versus time series of the campaign and the election), in country, in platform, and in question: the differential increase according to the candidate's local popularity, on one side, and the aggregate level of speech on American Twitter, on the other. A local contrast can detect what an aggregate series does not, and the reverse also happens; what accounts for the difference is something the data do not decide. The disagreement is part of the data, and the chapter does not resolve it.

DATA	In Brazil, hate speech in geolocated tweets increased after the 2018 election, more where Bolsonaro received fewer votes, concentrated in homophobia and sexism (Marino-Fages & Martínez-Pozo, 2026, working paper). In the United States, there was no persistent increase in hate speech on Twitter during the 2016 campaign or after the election, with the measure adopted (Siegel et al., 2021).
INTERPRETATION	The results are consistent with the idea that an election informs people about others' tolerance, and they diverge as to the aggregate effect; the measures, designs, and contexts differ too much for one to confirm or refute the other.
HYPOTHESIS	The information brought by the election reduced the sanction perceived by speakers who had previously held back. Neither study measures perceived sanction or the sanction that actually occurred.

10.5 Politicians and social media: who speaks from above and what they receive

If, as the Mackinac Center says (Section 10.1), politicians are more followers than leaders, the question is whom they follow. Frimer et al. (2023) offer the verified study that comes closest to answering, for American politicians on Twitter, and it speaks less of "the public" and more of the feedback the platform returns. Applying an incivility classifier to 1.3 million tweets from members of Congress since 2009, the authors find that incivility rose 23% over a decade (11.8 in 2009 versus 14.5 in 2019). The classifier, PerspectiveAPI, was built to assist forum moderation and gives each text a score between 0 and 100 that expresses how likely an observer would be to find it rude or disrespectful. The increase held in models that give each politician his or her own baseline level (an intercept per politician), so that it is attributable more to politicians who became more uncivil than to the entry of more uncivil politicians. In a pseudo-random sample of 998,935 U.S. tweets, incivility also rose among civilians (from 21.1 to 27.0, 28%), but the rise among politicians persisted with this control; it rose in the Senate and the House and across the political spectrum, with the largest increase among liberal Democrats, a difference the authors explain mainly by reactions to President Trump. In the presidents' tweets, mean incivility was 13.4 under Obama, with no change over time, and rose to 18.8 in 2017 and 23.0 in 2019 under Trump.

The part that matters for operant analysis is the feedback. Extremely uncivil tweets received on average ten times more retweets and eight times more likes than extremely civil ones, and the ratio grew: in 2009, uncivil tweets received about twice as many retweets; in 2019, almost 15 times as many (likes: 1.3 times and almost 8 times). The mediation analysis asks whether feedback explains the increase: when like feedback was statistically controlled, the increase in incivility fell by 74%; when retweet feedback was controlled, it disappeared. Because mediation is only correlational, the authors supplemented it with a Granger causality analysis within each account, which compares the moments when incivility was especially well received with the incivility that came afterward, and the analysis showed that periods of better reception were followed by more incivility. They conclude that politicians learned from the positive experience of tweeting uncivilly and therefore intensified it, and they attribute the increase both to the replacement, in Congress, of less uncivil politicians by more uncivil ones and to change in the politicians themselves.

Two results from the same study teach caution. The first: in floor speeches in Congress (1995 to 2020) and in presidential debates (1976 to 2016), incivility did not rise and may have fallen, which the authors take as support for the idea that social media may foster it. The second: likes and retweets do not imply approval. In supplementary studies, the more likes a tweet got, the less people really enjoyed it, and retweets of uncivil tweets seem driven by entertainment value, not by moral approval; the absolute count of likes is confounded by the greater exposure of uncivil tweets, and to correct for this the authors created the likes-per-retweet ratio. They also note that the sample includes 841 of the 1,003 members of Congress of the 111th to 116th Congresses, that the API limits collection to 3,240 tweets per account, and that the classifier's agreement with human observers was r = .627.

Read with the chapter's vocabulary, the study shows three things. The first is that what politicians emit on Twitter tracked the feedback the platform returns, and that statements at the edge (the uncivil ones) received more feedback than civil ones. The second is that the audience the politician "follows" is not "the public" but the subset that reacts, whose reaction, according to the authors themselves, diverges from approval. The third is that incivility did not rise in face-to-face arenas (the floor, debates), which is consistent with the idea that the window belongs to each arena, not to the politician, as the definition in Section 10.1 assumes. Chapter 9 (Section 9.6) already showed that, among members of the U.S. Congress, posts about the out-group were shared more in both camps (Rathje et al., 2021); the study by Frimer et al. (2023) shows an increase in incivility across the political spectrum, with the largest increase among liberal Democrats.

The study does not answer whether politicians moved the audience's window. It measures the politician's behavior and the feedback received, not what voters came to say. The three designs gathered so far each cover one side: Bursztyn et al. (2020) manipulate the information people have about local support; Marino-Fages and Martínez-Pozo (2026) observe users' speech after an election; Frimer et al. (2023) observe politicians' speech and the feedback. None links the politician's speech to the audience's with a design that identifies the direction. Asking whether the politician leads or follows, as the Mackinac Center page formulates the question, is asking something none of the three sources measures.

Finally, there is a methodological observation that serves as a warning. Kahmann and Heyer (2019) propose a natural language processing method to find, in newspaper texts, statements that shift the edges of the window. The idea is simple: in a reference period (8,067 documents from the newspaper *taz* on refugees, from 2013 and 2014), the words that usually appear alongside those

of the topic, the standard context, are mapped; then each sentence from 2015 receives a distance, from 0 to 1, which grows with the proportion of words that never appeared in that context, and only sentences with a statement indicator, such as verbs of saying and quotation marks, are kept, which reduces 14,894 sentences with a target word to 3,126. The authors themselves acknowledge that the use of new words does not always reflect an opinion outside the window, because many co-occurrences receive a high distance merely because they had not been used before, and this nonuse is due to random effects, not necessarily to being politically incorrect. They also acknowledge that the method does not guarantee that every sentence found is extreme: among those with the greatest distance is one about the educational success of migrants, which does not seem to exceed any window (a reading proposed here). What the method measures is novelty of context, not sanction; confusing the two is the error the chapter seeks to avoid.

DATA	In 1.3 million tweets from members of the U.S. Congress, incivility rose 23% in a decade, with much greater feedback (retweets and likes) for uncivil tweets, controlling for feedback reducing or eliminating the increase, correlational mediation and Granger analysis favoring reinforcement, part of the increase due to replacement of politicians, and no increase in floor speeches (Frimer et al., 2023).
INTERPRETATION	What politicians emit online varies with the feedback the platform returns, and statements at the edge were reinforced; the arena matters, because the same politicians did not increase their incivility on the floor.
HYPOTHESIS	Politicians moved their audiences' window, or followed it. No verified source measures audiences' speech in relation to politicians' speech with a design that identifies the direction.

10.6 Sanction, countercontrol, and what holds the edge in place

Chapter 5 dealt with hate speech as a disguised mand and as aversive control; this chapter takes from it only what serves the window: what makes the edge recede. The clearest verified study on this is the experiment by Álvarez-Benjumea and Winter (2018). Participants from a crowdsourcing platform (180, recruited on Clickworker, residents of Germany and paid three euros) commented, in a forum set up by the authors, on nine photos on four controversial topics (feminism, LGBT rights, refugees and multiculturalism, and poverty). They were required to comment on each discussion and saw only comments

chosen by the authors, drawn from a previous session in which three external raters classified 840 comments as friendly, neutral, or hostile. The discussions varied across four conditions: baseline (two comments each of the friendly, neutral, and hostile kinds), censored (the previous hateful content deleted), extremely censored (only friendly comments), and counter-speech (the hostile comments accompanied by replies that highlight the unacceptability of hate).

The participants' comments were rated by three judges blind to the conditions, on a scale from 1 (very friendly) to 9 (very hostile), with relatively high agreement among them (Krippendorff's alpha of .69, an agreement index on which 1 is perfect agreement); the mean of the three ratings formed the hostility score. Hostile comments (a rating of 7 or more) numbered 24 (5.88%) in the baseline, 4 (0.99%) in the censored condition, 18 (4.44%) in counter-speech, and 15 (4.10%) in the extremely censored condition; only the reduction in the censored condition was significant. Counter-speech had no significant effect, and the authors suggest as a possible explanation that the counter-comments, being themselves hostile, may add ambiguity about the norm by setting descriptive and injunctive information in conflict. They note that they have no data on the trade-off between censorship and freedom of expression and that the anonymity of the forum precludes group identification processes; the study takes no position on censoring. Chapter 5 (Section 5.5) takes up, with the same data, the authors' conclusion that normative behavior online is driven more by descriptive norms (what others do) than by injunctive norms (what one ought to do).

In Skinner's (1957) terms, counter-speech is a sanction without power: it came from anonymous peers, with no control over the speaker's reinforcers, and nothing in the design turns it into an “effective” negative audience. Far from being a definitive finding, this is an analytic reading derived from the authors' explanation (according to which the reply itself carries hostility), since the research did not manipulate the authority of the sanctioning source. The result matters, however, for practice: in the experiment's data, hostility did not decrease because someone said it should not occur; it decreased when the environment displayed less of it. The same rule should hold in the opposite direction, so that an environment that comes to display more of what was previously outside the window would widen it; this is a hypothesis proposed here, and the experiment did not test it.

Countercontrol (Delprato, 2002) enters by the same path. The sanction that punishes a statement is a socially mediated aversive contingency; those who suffer it may escape or avoid it, and one of the four avoidance repertoires described in Section 6.4 is that of place, migration to closed, unmoderated

spaces. In Skinner's (1957) terms, this is seeking the confidant, the nonpunishing audience. Hence the hypothesis, proposed here: the bubble of Chapter 6 is a collective nonpunishing audience for certain topics, which widens the speaker's effective window while he or she remains in it, and Skinner's dilemma appears when a post reaches, at the same time, the bubble and an audience with a narrower window. Skinner (1957) says that, as long as the audiences remain separate, the dilemma does not arise. If platforms bring together in a single emission audiences that were previously separate, the dilemma arises more often; no verified source measures this.

TABLE 10.1

What each study manipulated or observed, what it measured as "saying," and what it does not allow us to conclude

Study	Material and design	What it measured as "saying"	What was observed	Caveat
Bursztyn et al. (2020)	Two online experiments in the U.S., 2018: 1,600 Pittsburgh residents; about 1,800 Democrats. Randomly assigned information about who won the election in the area	Authorizing a donation to an anti-immigration organization, in private and in public (action); in Exp. 2, amount given to the donor	In public, 33% (Trump won) and 20% (Clinton won); in private, 31% and 30%. Observers gave more to a public donor where Trump won	Manipulates information, not a politician's statement; action, not statement; online panel; observers only Democrats
Marino-Fages and Martínez-Pozo (2026)	Brazil's 2018 election as a surprise; difference-in-differences by municipality; more than 37.6 million geolocated tweets	Hate speech by BERT classifier (F1 of 77 to 79%) and by dictionary	Increase after the election, larger where Bolsonaro received fewer votes; concentrated in homophobia and sexism	Working paper, not peer-reviewed; only 4 to 8% geolocated; parallel trends assumed
Siegel et al. (2021)	Series of more than 750 million tweets from the 2016 election and almost 400 million from a random sample	Dictionary augmented by machine learning and classification by Reddit communities	No persistent increase in hate speech during the campaign or after the election; brief spikes	Dictionary requires a slur; Twitter only; authors point to other platforms
Frimer et al. (2023)	1.3 million tweets from members of the U.S. Congress, 2009–2019; floor speeches and debates	Incivility by classifier (PerspectiveAPI; r = .627 with observers)	+23% in a decade; greater feedback for uncivil tweets; no increase on the floor	Accepted manuscript; incivility is not hate speech; correlational mediation; partly due to replacement

Study	Material and design	What it measured as "saying"	What was observed	Caveat
Álvarez-Benjumea and Winter (2018)	Online experiment with 180 crowdsourcing participants; forum with four conditions	Comments with a hostility rating of 7 or higher	5.88% at baseline; 0.99% censored; 4.44% counter-speech; 4.10% extremely censored	Paid, anonymous participants; no data on the trade-off with freedom of expression
Kahmann and Heyer (2019)	Method applied to a German newspaper (taz); 8,067 reference documents (2013 and 2014) and texts from 2015	Novelty of context of each sentence	List of sentences with the greatest distance	Novelty is not impermissibility; the authors admit that not every sentence found is extreme

Note. The figures come from the articles. The page numbers of Bursztyn et al. (2020) are those of the American Economic Review, and those of Kahmann and Heyer (2019), those of the proceedings; the other sources have no journal pagination in the versions read. The column "What it measured as 'saying'" describes what each study counted; none measured the sanction that the saying received, and only Álvarez-Benjumea and Winter (2018) varied it.

DATA	In an experimental forum, removing previous hateful content reduced hostile comments from 5.88% to 0.99%, and adding replies pointing out its unacceptability had no significant effect (Álvarez-Benjumea & Winter, 2018).
INTERPRETATION	What the environment displays as done by others weighs more, in these data, than the verbal reply; a reply from peers without power does not have the status of an effective negative audience (a reading proposed here, based on Skinner, 1957).
HYPOTHESIS	The bubble is a collective nonpunishing audience for certain topics, and the platform, by bringing audiences together, multiplies dilemmas. No verified study tests these.

10.7 How to situate a statement in relation to the window

The chapter does not offer a criterion for saying that a statement is "outside the window": that would require measuring the sanction, and the sources do not measure it. What can be offered is a set of questions that prevents the most common errors of everyday conversation. The instrument is for the analyst, and it does not classify statements.

Questions 1 to 4 come from the definitions in Sections 10.1 and 10.2; questions 5 to 8, from the limits of the studies gathered in Table 10.1. Question 7 repeats the criterion of the entire book: the same method for all camps.

Questions for situating a statement in relation to the window

1. What is the audience, and in what arena did the statement occur (legislative floor, social network, radio, family)? The same person may have different windows in different arenas.
2. What is known, from records (and not from impressions), about the sanction that similar statements received from this audience? Without records, the analyst should say that he or she does not know.
3. Are audiences with different windows present in the same emission (dilemma)? The platform decides which ones.
4. What changed between two moments: who is listening, the power of sanction, information about others' tolerance, what the environment displays, what is reinforced at the edge, repetition, or reach?
5. What was measured as "saying": raw text, classifier, dictionary, action? With what error, and with what definition of "hate" or "incivility"?
6. Against what was the comparison made: the same speaker before and after, or other speakers, and with what selection criterion?
7. Was the same criterion applied to statements from opposing political camps?
8. What does the measure fail to distinguish (novelty, incivility, hate, action)?

10.8 Rival interpretations and limits

The chapter's reading, that an audience's window is the set of statements it does not punish and that the edge shifts when the sanction changes, has rivals. The box below gathers them, one by one, and the chapter's limits come after it.

Rival interpretations for the same set of data

A change in composition, not in the window. Frimer et al. (2023) attribute the increase both to the replacement of less uncivil politicians by more uncivil ones and to change in the politicians themselves; what is seen in the aggregate is the sum of the two, and only the second is a change in the speech of those who were already speaking.

A change in measurement. The studies use classifiers with an F1 of 77 to 79% (Marino-Fages & Martínez-Pozo, 2026), dictionaries that require a slur (Siegel et al., 2021), and an incivility classifier with r = .627 with observers (Frimer et al., 2023); "hate speech" and "incivility" are not the same construct.

Other channels, not the norm. Bursztyn et al. (2020) say that the treatment effect may act through other channels, although they consider these unlikely; the reading in terms of norms is the authors'.

An event, not a window. Marino-Fages and Martínez-Pozo (2026) exclude the months from August to October 2018 and assume that, without the election, trends would have been parallel across municipalities; the assumption may fail, and other events of the period may have acted differentially across municipalities (an observation made here).

An unfalsifiable metaphor. If any change in speech counts as a shift of the window, the concept does not separate hypotheses. That is why it is defined here by the audience's sanction (Section 10.1), and a measure of the sanction is called for (Section 10.7).

Direction and symmetry. The election studies deal with the victories of Trump and Bolsonaro; none deals with the mirror case. The concept is symmetrical by definition, and the data closest to symmetry, in the sense that the increase was general, come from Frimer et al. (2023), in which the largest increase occurred among liberal Democrats, explained by the authors mainly by reactions to Trump; the symmetry of an election's effect has not been tested.

An interested source. The description of the concept comes from a think tank that uses it to promote its own policies (Mackinac Center, 2019).

Sample and platform. An online panel from one region and Democratic observers (Bursztyn et al., 2020), paid crowdsourcing participants (Álvarez-Benjumea & Winter, 2018), Twitter (the others). Nothing here speaks to WhatsApp, YouTube, or TikTok.

Five limits apply to the entire chapter. First, the concept of the window, in its origin, is about public policy and comes from an interested party; the translation into statements and audiences is proposed here, and no source tests it as such. Second, the measures differ (perception, public action, classified text) and do not add up; the comparisons in Table 10.1 are of order of magnitude and design, not of effects. Third, no source measures the sanction, which is the central term of the definition: the experiment by Bursztyn et al. (2020) touches on the observer's side, and that of Álvarez-Benjumea and Winter (2018) varies what the environment displays, but neither measures the punishment the speaker actually suffers. Fourth, the Brazilian source is a working paper about one platform and one election, and none measures the speech of Brazilian politicians; that is the question one would like to ask, and the verified sources do not answer it. Fifth, the data that would link the politician's speech to the audience's, with identification of direction, probably require records that only the platforms have, as in the limits of Chapter 8.

For those who want to use the chapter in the analysis of statements (Chapters 3, 5, and 7), the consequence is practical. To say that a statement "shifted the window" is to assert three things, all empirical: that the audience would have punished it before, that it stopped punishing it, and that the politician's statement played a part in this. The analyst who asserts only the first describes an impression. Chapter 11 takes up, when dealing with measures of diffusion, the distinction among reach, exposure, and reaction, which matters here because the effective window of a post depends on who sees it; Chapter 12 (Section 12.6) deals with interventions on the platform's contingencies, and the criterion used here, the same for what widens and for what narrows the window, applies to them.

Summary

The Overton window, in its origin, is a model of which policies a politician can support without losing popular support, formulated by a think tank that uses it to promote its own policies and that asserts that politicians are more followers than leaders. Here it is translated into the analysis of verbal behavior: an audience's window, for a given speaker and topic, is the set of statements that this audience does not punish; it is a relation, not a property of the statement; there are as many windows as there are audiences, and Skinner's dilemma appears when two audiences with different windows are present (Sections 10.1 and 10.2).

The verified data show different pieces of the phenomenon. Information about who won a local election altered what Pittsburgh residents agreed to do in public, without altering what they did in private, and altered what observers gave to those who acted in public (10.3). A surprising Brazilian election was followed by a differential increase in hate speech in tweets, larger where the candidate received fewer votes, in a working paper; an American series found no persistent increase on Twitter, with a different measure (10.4). Members of the U.S. Congress became more uncivil on Twitter to the extent that incivility received more feedback, but not on the floor, with part of the increase due to replacement of politicians and with the caveat that likes are not approval (10.5). In an experimental forum, removing hate reduced hate, and replies condemning it did not (10.6).

What the data do not show is the central term and the direction that the underlying question calls for. No study measures the sanction, or links the politician's speech to the audience's with identification of direction, and none deals with a mirror case. In the sources, the window is the name of a result (the change in what is done or said in public), not of a mechanism; in Section 10.7,

the chapter offers questions to help the analyst distinguish what he or she knows from what he or she assumes. Chapters 3, 5, and 7 use the idea that the window is specific to each audience and each arena; Chapter 6, the idea that the bubble is a nonpunishing audience (a hypothesis); Chapter 8, the idea that reach decides how many people hear; Chapter 9, the idea that the politician is a producer who responds to feedback; Chapter 12, the idea that intervening means changing the contingencies that hold the edge in place.

References

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, *110*(11), 3522–3548. https://doi.org/10.1257/aer.20171175

Delprato, D. J. (2002). Countercontrol in behavior analysis. *The Behavior Analyst*, *25*(2), 191–200. https://doi.org/10.1007/bf03392057

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, *14*(2), 259–269. https://doi.org/10.1177/19485506221083811

Kahmann, C., & Heyer, G. (2019). Measuring context change to detect statements violating the Overton window. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 392–396. https://doi.org/10.5220/0008191803920396

Mackinac Center for Public Policy. (2019). *The Overton window*. https://www.mackinac.org/OvertonWindow

Marino-Fages, D., & Martínez-Pozo, A. (2026). *Hate in the Tropics: Political leaders and the social acceptability of online hate speech* (Documento de trabajo N.º 391). RedNIE. https://rednie.eco.unc.edu.ar/files/DT/391.pdf

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 U.S. election campaign and its aftermath. *Quarterly Journal of Political Science*, *16*(1), 71–104. https://doi.org/10.1561/100.00019045

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

instituto WALDEN4 | ACADEMIA WALDEN4



RÁDIO IW4

Articles, lectures, and classes to listen to

Readings of scientific articles, lectures, and classes in audio, 24 hours a day, 7 days a week: for studying on the way, in traffic, or at the gym.

24h a day, 7 days a week

Free access, no sign-up required

walden4.com.br/radio

CHAPTER

11

Infodemiology: How Information Spreads and What Is Measured

Cinelli et al. (2020), in a study of the COVID-19 infodemic, open their article with a case. According to them, a news outlet anticipated the rumor that Lombardy, in northern Italy, would be locked down and published it hours ahead of the Italian Prime Minister's official announcement; in consequence, people crowded trains and airports to go south before the lockdown, which, in the authors' words, hampered the government's initiative to contain the epidemic and possibly increased contagion. The case is reported as the authors tell it; the news item was not read, and it is not known how people's behavior was measured.

The case serves the chapter for a reason that is not the authors'. The information that circulated is not presented as false: it was a rumor about a measure that was in fact taken, and the harm attributed to it comes from when and how people acted on it. Public health distinguishes rumor from disinformation for precisely this reason, because a rumor may turn out to be true (Calleja et al., 2021). This reading is proposed here, and the chapter's first lesson follows from it: what concerns those who study an infodemic is not only false information but the volume, speed, and quality of information, and the veracity of each item is one variable among others that are important to consider.

The chapter asks how the diffusion of information is measured, what the verified data show about the true and the false, and how this connects to the contingencies of the previous chapters. Section 11.1 presents infodemiology and the word infodemic; Section 11.2 deals with measures of diffusion and the distinction between reach, exposure, and reaction; Section 11.3 examines the finding that falsehood spreads faster than truth; Section 11.4, epidemiological

models applied to information; Section 11.5, the COVID-19 infodemic, worldwide and in Brazil; Section 11.6, what the field proposes to do; and Section 11.7 brings together questions for reading a diffusion study.

Three cautions guide the reading. The first is that epidemiological language is an analogy. No infectious agent passes from one person to another; what spreads are statements, and each passage is someone's verbal response (Chapter 3), maintained by consequences that the contagion model has no way to represent. The analogy is useful for measuring and risky for explaining. The second is that measures of diffusion count responses of passing something on, not belief, nor effects on what people do afterward. The third is coverage: almost everything available comes from Twitter, in English, and from the beginning of the pandemic; for Brazil, the chapter has a single study of links about vaccines.

11.1 Infodemiology and infoveillance

The term infodemiology is Eysenbach's, who defines it as the science of how information is distributed in an electronic medium and of what determines that distribution, with the aim of informing public health and public policy (Eysenbach, 2009). The name is a portmanteau of information and epidemiology, and the analogy is literal in the proposal: just as epidemiology counts cases and looks for determinants, infodemiology counts occurrences of information and looks for what determines them. The 2009 article distinguishes two kinds of indicators: supply-based ones, which count what is published (web pages, messages, blogs), and demand-based ones, which count what people search for and where they navigate. The basic metrics are information prevalence, how many times, in absolute or relative terms, a keyword or concept occurs in a pool of information; incidence, how many new units of information appear per unit of time; and the occurrence ratio, which compares a concept with a control term, because the total size of the pool is often unknown. Using these data for surveillance is what the author calls infoveillance, and the methods that analyze what is available without involving people he calls passive.

The founding example comes from health, not from disinformation. A 2006 study by Eysenbach himself showed that Google searches about influenza correlated with influenza cases in Canada the following week, and that the searches preceded visits to sentinel physicians by one week (Eysenbach, 2009). Because Google did not make its search data available to outside researchers, the study purchased a keyword-triggered ad, which returned search and click statistics. The detail matters for two reasons. It shows that

measures of diffusion depend on data that the platforms control, and Lazer et al. (2018) say that the cooperation of platforms with researchers will be critical for understanding the scale of the problem and the effectiveness of interventions. It also shows what these data are: records of what people do, measured as searching and clicking, not of what they say, which the author presents as an advantage over surveys, since they are direct and free from recall or social desirability biases (Eysenbach, 2009).

The advantage comes at a cost that the author himself records. A search in isolation is ambiguous as to the searcher's intention (someone searching for "cold" does not necessarily have the symptoms), and search engines' hit counts are neither accurate nor reliable (Eysenbach, 2009). In the language of Chapter 3, the form of the response, the word searched, does not tell its function. The author also acknowledges the arrow in the other direction: changes in information patterns can affect population health, as in an "outbreak" of misinformation or in a campaign, and in 2009 he writes that no robust methods yet existed for automatically detecting and monitoring how far information that lowers vaccination rates had spread (Eysenbach, 2009). The sentence dates from 2009, and no claim is made that it is still true; it marks the field's starting point.

The word infodemic became widely known in 2020. On February 15 of that year, the Director-General of the World Health Organization (WHO) warned that the epidemic was being accompanied by an infodemic (Calleja et al., 2021), and the organization defines it as an overflow of information of varying quality, in digital and physical environments, during an acute health event, which leads to confusion, risk-taking behaviors, and loss of trust in health authorities. Eysenbach (2020) writes that he coined infodemiology in 2002 and, over twenty years, infodemic and infoveillance; the 2009 article, which was one of the sources consulted, uses the terms infodemiology and infoveillance but does not contain the word infodemic.

Zielinski (2021) reduces the primary characteristics of an infodemic to two, the volume of information generated and the velocity with which it appears, and treats the forms of wrong information as secondary characteristics, albeit the most significant ones. The author asks whether a large volume of largely true and useful information, appearing quickly, would be an infodemic, and answers that, for an information manager, any sudden, large, fast-moving surge of information calls for management. The article's table associates with volume the problems of locating, storing, assessing quality, giving visibility, and validating, and with velocity, the problems of assessing, filtering, applying, tracking history, and avoiding waste, such as research unnecessarily replicated. The examples come mainly from the scientific literature: the increase

of over 400% in preprint postings on medRxiv (from 586 in the last 15 weeks of 2019 to 2,572 in the first 15 weeks of 2020), which the article cites from another work.

This reorganizes the question. Chapters 1 to 7 dealt with false statements and what maintains them. Infodemiology measures a phenomenon of a different order: the volume and speed of information, true or false, and what this does to the recipient's ability to tell one from the other. Falsehood is part of the problem, not its definition. Eysenbach (2020) adds a caution that the chapter adopts. At the beginning of a pandemic, "facts" are better called "BETs" (best evidence at the time); recommendations based on them are subject to change, and spreading facts to counter the infodemic is easier to propose than to carry out when nobody is sure what the facts are. The author gives as examples the recommendations on masks, the use of certain drugs, and distancing rules. The consequence for any measure of "false information" in health is proposed here: the label of false depends on who checks, with what evidence, and at what moment, and Chapter 2 dealt with this dependence for fact-checking agencies.

11.2 Measures of diffusion: reach, exposure, reaction, and cascade structure

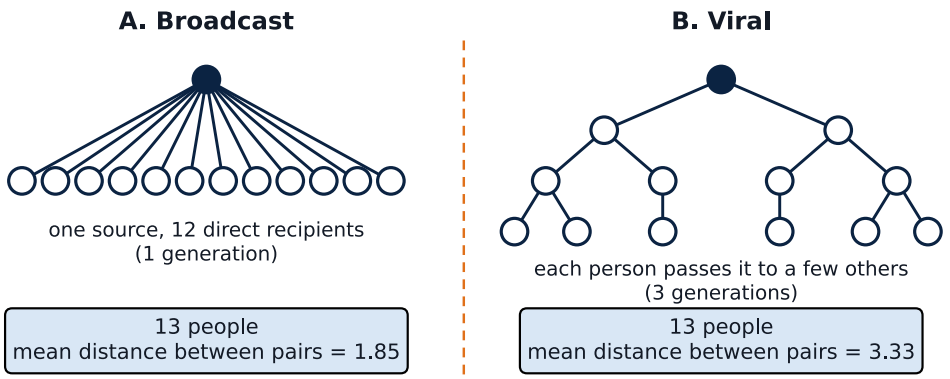
Measuring diffusion requires deciding what counts as a person who "received" something and what counts as one who "passed it on." The studies in this chapter use the cascade as the unit: the set of reshares linked, by an unbroken chain, to a common origin. A cascade can be of size one, if no one reshared the original emission, and a rumor can have several cascades, as many as the times it was introduced independently (Vosoughi et al., 2018). Each cascade is described by four numbers: size, the number of users involved; depth, how many retweet hops separate a point in the cascade from its origin, each hop being a retweet by a user not yet in the cascade; maximum breadth, the largest number of users involved at any depth; and structural virality.

Structural virality is used by Goel et al. (2016) and responds to a problem of interpretation. When something is said to have "gone viral," it is understood to have grown from person to person, like a virus; but content can reach millions through a single large broadcast, such as one from an account with tens of millions of followers, without anyone passing it on to anyone else. Final size does not distinguish the two cases. Depth also fails, because a single long chain within an enormous broadcast yields high depth. The authors define structural virality as the average distance between all pairs of people in the cascade tree, a measure known in mathematical chemistry as the Wiener in-

dex, which does not depend on assumptions about the process that generated the diffusion. In a pure broadcast it is close to 2, the minimum; the more generations, the higher it is. Figure 11.1 shows two trees of 13 people and the difference the number records.

FIGURE 11.1

Broadcast and viral diffusion: two trees of the same size



Same size, opposite structures. Original diagram, based on the concept in Goel et al. (2016); the mean distance was calculated for the trees as drawn. Black circle: the person who introduced the content.

Note. Original diagram, based on the concept in Goel et al. (2016); without data. The mean distance between all pairs was calculated for the two trees as drawn; in a broadcast with many people, the value tends toward 2.

The dataset of Goel et al. covers about a billion diffusion events on Twitter between July 2011 and June 2012: tweets with links to popular news sites, videos, images, and petitions. What they found dispels the image of the viral. The mean tree size is 1.3, and about 99% of adoptions are by the roots or their immediate followers; only 0.025% of trees reach 100 nodes or more, 219,855 out of about 1 billion. Among these there are almost all structures, from pure broadcast to branching ones, but median structural virality is low and varies little with size: less than 3 for images and videos, even with 10,000 reposts or more; between 7 and 8 for petitions; and below 8 for the largest news cascades, of 3,000 reposts. The correlation between size and structural virality is 0.2 for news, 0.04 for petitions, and essentially zero for images and videos, and the authors conclude that, insofar as popularity tracks any single feature consistently, that feature is the size of the largest broadcast. As to causes, they say they can only speculate: they suspect a lack of large broadcast channels for petitions and, for images and videos, celebrities with tens of millions of followers.

The result holds within the limit that the authors themselves set: it is restricted to the diffusion of information on Twitter, and without reconstructing the sequence of events and without a measure of virality, popularity alone says little about how much the viral and the broadcast mechanisms each contrib-

ute. The reconstruction is also an inference: each adopter is linked to the friend, that is, the followed account, who adopted the content most recently before them, and the authors say they correctly infer the parent in about 95% of cases.

Three terms organize what the measures count, and they are fixed here because Chapters 8 and 10 use them and the literature mixes them up.

PROPOSED DEFINITION

Reach, exposure, and reaction.

Reach is the set of people to whom an emission is delivered or who can see it; it depends on the network of contacts and on the platform's delivery rule (Chapter 8). *Exposure* is what a person actually saw. *Reaction* is what one does after seeing: liking, commenting, resharing, or ignoring. Skinner does not use these terms; the distinction is proposed here and serves to say, of each measure, which of the three it counts.

The studies in this chapter count mainly reaction, and when they speak of exposure they infer it. Goel et al. presume exposure from the list of followed accounts; Mosleh and Rand (2022) build each user's score of exposure to lying elites from whom the user follows, and point out that exposure and sharing are constructs of a fundamentally different kind: most people share just a small part of what they see. Reach, in turn, is what the platform sets, and Chapter 8 showed that ranking alters the reach of each emission unequally (Huszár et al., 2022). Table 11.1 states, for each study in the chapter, what counts as diffusion and what is left out.

None of these measures says what the emission did to the behavior of those who saw it. Lazer et al. (2018) write that exposure does not equal impact and that studies evaluating how exposure to fake news affects political behavior over the medium and long term are essentially absent from the literature. The chapter returns to this point in Section 11.6.

TABLE 11.1

What each study counts as diffusion and what is left out

Study	What it counts	Unit	Measures	Does not measure
Eysenbach (2009)	Searches (demand) and publications (supply) with a keyword or concept	Term or concept in a pool of pages or searches	Occurrences and searches	Who saw it; what they did with the information; the intention of the search
Goel et al. (2016)	Reposts of links, linked to the followed account that reposted earlier	Tree of each piece of content	Reaction (reposting); exposure is presumed	What the platform delivered; who saw without reposting; the veracity of the content
Vosoughi et al. (2018)	Retweets of a fact-checked claim	Rumor cascade	Reaction (retweeting), in number of users	Who saw without retweeting; which rumors no one checked
Cinelli et al. (2020)	People who post about the topic; posts and reactions by class of outlet	Authors per day; posts and reactions per link	Reaction (posting, commenting, reacting)	Exposure; the veracity of each claim (the label belongs to the outlet)
Massarani et al. (2021)	Total engagement with the link (likes, comments, shares, reactions)	Link among the hundred with the highest engagement	Aggregate reaction, without distinguishing type	Reach; links outside the hundred; who reacted

Note. Our own elaboration, based on the articles' methods. The "Measures" column uses the three terms of the definition above.

11.3 Does falsehood spread faster than truth?

The field's most cited answer to this question is that of Vosoughi et al. (2018), and it is worth knowing the design before the result. The authors gathered the cascades of rumors that six fact-checking organizations had verified and that circulated on Twitter from 2006 to 2017: about 126,000 stories, tweeted by about 3 million people more than 4.5 million times, classified as true or false with 95 to 98% agreement among the organizations. Data collection started from replies in English to tweets carrying a link to one of the fact-checking

pages; the original tweet and its retweets form the cascade. The authors avoid the expression "fake news," considering that it has become politically polarized and unrelated to veracity. Instead, they use "true" and "false" and say they make no claims about the intent of those who spread the information.

Falsehood spread significantly farther, faster, deeper, and more broadly than truth did, across every category of information, and the effect was stronger for political falsehoods than for those about terrorism, natural disasters, science, urban legends, or finance. The truth rarely reached more than 1,000 people, whereas the top 1% of false cascades routinely reached between 1,000 and 100,000; moreover, true stories took roughly six times longer than false ones to reach 1,500 people. In a model of the probability of retweeting, falsehoods were 70% more likely to be retweeted, with controls for the originating account's age, activity, numbers of followers and followees, and verified status. One datum calls for care: a greater fraction of false rumors had between 1 and 1,000 cascades, and a greater fraction of true rumors had more than 1,000. The truth, therefore, more often had many independent origins, and the result for falsehood is a result about reach per cascade.

The authors propose an explanation and present it with caution. Across all novelty metrics, calculated against the tweets users had been exposed to in the previous 60 days, falsehood was more novel than the truth, although users may not have perceived it as novel. In the replies, according to an emotion lexicon, falsehood inspired more surprise and disgust, and the truth more sadness, anticipation, joy, and trust. But they write that they cannot claim that novelty causes retweets or that it is the only reason falsehood is retweeted more, and that the behavioral explanation requires research involving direct contact with users. Regarding bots, the authors found that they sped up the spread of true and false news equally, and they conclude that falsehood spreads more because humans, not bots, are more likely to spread it.

The limits are those of the design. Only rumors that someone checked are included, and the organizations may select certain kinds of rumors or draw attention to them; the authors tested this objection with 13,240 cascades that students, working independently, investigated and labeled as true, false, or mixed, with 90% agreement and a Fleiss' kappa of 0.88 (an index of interrater agreement that discounts the agreement expected by chance), and found nearly identical results. Everything is Twitter, in English, up to 2017. And "farther" means more new users who retweeted, not more people convinced. There is also an apparent tension with the study by Goel et al. in the previous section. Vosoughi et al. say that falsehood did not simply spread through broadcast, but through peer-to-peer diffusion, in a viral branching process; Goel et al. show that, on average, large cascades are dominated by broadcast.

The two fit together, because one compares false and true and the other describes the whole set, and they are a reminder that "viral" is a structural measure, to be compared, not a label to be applied to everything that is large.

Besides Vosoughi et al., two other studies in the chapter ask the same question with a different unit, and Table 11.2 sets them side by side.

TABLE 11.2

Three studies, three units: what each compared between the reliable and the false

Study	Unit compared	Who classified	Platform and period	What was found
Vosoughi et al. (2018)	Fact-checked claim (rumor cascade)	Six fact-checking organizations	Twitter, 2006–2017	Falsehood reached farther, faster, deeper, and more broadly; the effect was greater for political content
Cinelli et al. (2020)	Link, labeled by outlet (reliable or questionable)	Media Bias/Fact Check	Five platforms, January 1 to February 14, 2020, on COVID-19	Posts and reactions from reliable and questionable sources grow at the same pace within each platform; relative amplification varies with the platform
Massarani et al. (2021)	Vaccine link, among the hundred with the highest engagement	The authors themselves, with categories adapted from another work	Facebook, Twitter, Pinterest, and Reddit, in Portuguese, 2018–2019 and 2020	Equal proportion of disinformation in both periods; higher mean engagement for disinformation only in 2020

Note. The results come from the articles; the units and criteria are in Sections 11.2 and 11.5. These measures are not comparable with one another.

No work reexamining the comparison of Vosoughi et al. with another method was read. Three works planned for this chapter could not be obtained by legal, automated means (the list is in Appendix D). The chapter therefore does not say what they conclude, and the result of Vosoughi et al. appears here without whatever scrutiny later work may have given it.

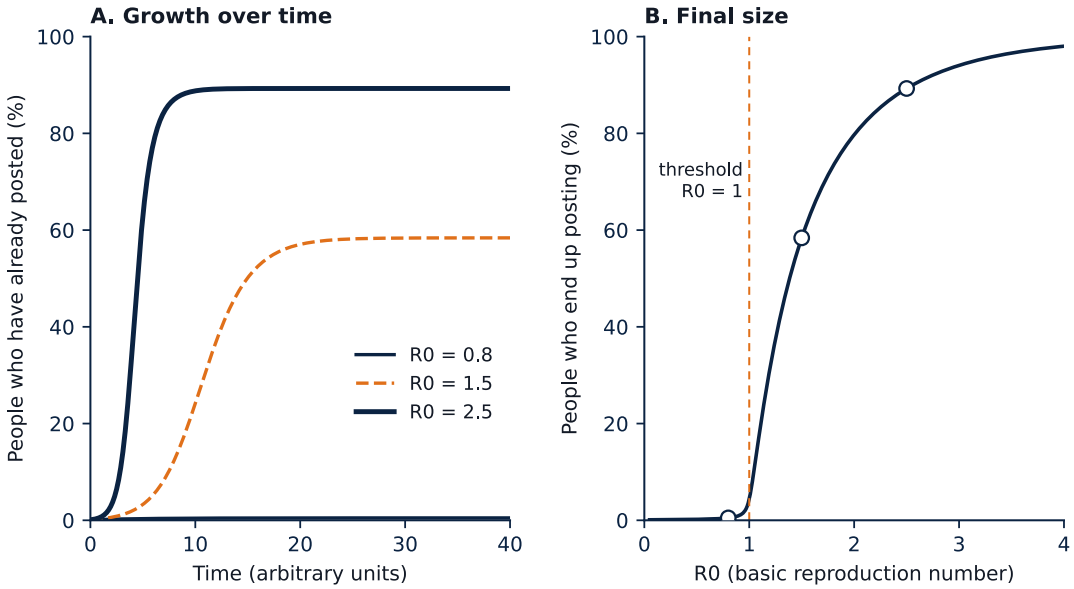
DATA	Fact-checked false claims reached farther and faster than true ones on Twitter, from 2006 to 2017 (Vosoughi et al., 2018). On five platforms and in the first 45 days of 2020, links from outlets labeled questionable and reliable grew at the same pace within each platform, and relative amplification varied among them (Cinelli et al., 2020). In 94 and 89 vaccine links, the proportion of disinformation was practically the same in 2018–2019 and 2020, and the mean engagement of disinformation exceeded that of verified content only in 2020 (Massarani et al., 2021).
INTERPRETATION	"Falsehood spreads more" is a claim about one unit (the fact-checked claim), one platform, and one period. The same question with another unit, the outlet, or on another platform, does not yield the same result, and none of the studies contradicts the others, because they measure different things.
HYPOTHESIS	When falsehood spreads more, it is because content properties, such as novelty, control resharing responses that the audience already emitted, in the sense of Chapters 8 and 9. This is compatible with the novelty, surprise, and disgust data of Vosoughi et al., which have no causal test; the authors themselves say they cannot claim that novelty causes retweets.

11.4 Epidemiological models applied to information

Epidemiological models divide the population into compartments, susceptible, infected, and recovered (SIR); an infected person infects each susceptible contact with a certain probability; and the basic reproduction number, R0, is the average number of secondary cases that one infected person produces. In epidemiology, R0 = 1 is the threshold: below it the disease dies out, above it the disease spreads (Cinelli et al., 2020). Goel et al. (2016) write R0 as the product of the average number of contacts of each node and the probability of contagion per contact. Figure 11.2 shows what the threshold does, in our own simulation, with illustrative parameters and no data.

FIGURE 11.2

The R0 threshold in a simple SIR model



Original simulation: homogeneous-mixing SIR model (everyone has the same contact with everyone), with 0.1% initial posters and recovery of 1 per unit of time. Illustrative parameters; no data from any platform, topic, or period.

Note. Original simulation: homogeneous-mixing SIR model (everyone has the same contact with everyone), with 0.1% of people posting at the start and recovery of 1 per unit of time. Illustrative parameters; no data from any platform, topic, or period. Panel B: final size calculated for each R0.

Cinelli et al. (2020) applied the model to 1,342,103 posts and 7,465,721 comments by 3,734,815 users on five platforms, from January 1 to February 14, 2020. What grows is the number of authors posting about the topic, and the "infected" person is one who starts posting after being exposed to it; in the authors' approach, R0 greater than 1 would indicate the emergence of an infodemic. They fitted two models: a phenomenological one (EXP), which seeks to reproduce the data without assuming a specific growth mechanism, and the standard SIR (*Susceptible*, *Infected*, *Recovered*) model. All estimated values are above 1 (they are "supercritical"), even taking confidence intervals into account: with EXP, from 1.42 to 1.52 on Gab[1] and from 2.02 to 2.64 on Instagram; with SIR, from 2.2 to 2.5 on Gab and from 4.0 to 5.1 on Twitter, and about 100 on Instagram, a value that the authors say is far above anything observed in a real epidemic and that they attribute to an abrupt jump in new authors that continuous models do not explain.

[1] An American microblogging network.

Two cautions apply to these numbers. The first concerns what R0 is here: a parameter of growth in the number of people posting about a subject on a platform, over 45 days, and not a property of a false claim or of a person. The

second is that the fit of the model to the curve does not establish the mechanism. Goel et al. (2016) write that reproducing certain empirical results with a theoretical model does not, by itself, mean that the model faithfully captures the process that actually generated the data, although it is possible to rule some models out.

The most instructive use of the model, in the literature consulted, is that of Goel et al., who employed it to rule out hypotheses. The random-network model is ruled out because it would produce sizes concentrated around the mean, not the heavy tail; with a scale-free network and low infectiousness (r around 0.5, that is, subcritical), the model reproduces the probability that an emission becomes popular, the mean structural virality, and the correlation with size, but not the observed diversity of structures. The authors conclude that the popular *tipping point* notion has, for the most part, little bearing on the viral events they study, which runs against the common idea that information "goes viral" by passing a threshold: everything they observed, including the largest events, fits a low-infectiousness model in which, in a network with a few hubs with many followers, a piece of content easily finds a node that sustains it.

In behavioral language, what the model calls infection is a verbal response: posting or passing on a message in the presence of another person's emission. The model treats this response as an event that exposure triggers with equal probability for all susceptible people and has no term for the consequence that maintains it. This is what Chapters 6, 8, and 9 opened up: diffusion is reinforced by its effects on the reactions of a specific audience, and the probability of diffusing varies with the audience (Chapter 6), the content (Chapter 8), and the history of the person passing it on (Chapter 9).

CONCEPT IN USE

Macrocontingency (Glenn, 2004), revisited from Chapter 9

Chapter 9 (Section 9.7) distinguished the metacontingency from the macrocontingency. A macrocontingency is the relation between a cultural practice and the summed consequences of the macrobehavior that makes up that practice; these cumulative effects, Glenn (2004) writes, cannot control, as operant consequences, the behavior of any individual. A cascade of 10,000 reshares is an aggregate effect of this kind: none of the 10,000 was maintained by the size of the cascade. What maintains each reshare is the feedback the person receives from their audience. For Glenn, effective cultural engineering requires identifying the macrocontingencies that produce undesirable effects and changing the operant contingencies or metacontingencies involved. Reading proposed here: reducing the size of cascades calls for changing what maintains each person's resharing, not the size itself.

This does not make scale irrelevant. What concerns public health is the aggregate effect, and that is what the epidemiological model measures. The distinction is one of level: the phenomenon is measured at the level of the population and is explained, in the reading proposed here, by the interlocking of different people's contingencies.

DATA	SIR models fit the growth curve of the number of authors on a topic on five platforms (Cinelli et al., 2020), and a SIR model on a scale-free network with low infectiousness reproduces the size distribution of cascades on Twitter, but not the diversity of their structures (Goel et al., 2016).
INTERPRETATION	The model describes growth without describing what maintains each resharing response; its contagion parameter combines reach and probability of reaction into a single number.
HYPOTHESIS	The probability of resharing is a product of the contingencies of Chapter 9 (audience reaction, outrage, group membership) and varies across people and content in a way that the homogeneous-mixing model does not capture. This is compatible with the model's failure to reproduce structural diversity, but that failure has other explanations, and no source read tests the hypothesis.

11.5 The COVID-19 infodemic, worldwide and in Brazil

Two studies on the COVID-19 infodemic, one on a global scale and one on Brazil, make it possible to advance our understanding of this process. The study by Cinelli et al. (2020) compares COVID-19 in five environments. Twitter, Instagram, and YouTube are mainstream platforms; the other two are Gab, which the authors describe as crowdfunded, with very little control over content and a user base considered far-right, and Reddit, a news aggregator based on collective filtering. The authors classified links as reliable or questionable according to Media Bias/Fact Check, which labels outlets rather than articles, and they make this limitation explicit. Within each platform, posts and reactions from questionable and reliable sources grow at the same pace, and the scaling factor depends strongly on the platform. On Gab, posts from questionable sources were about 70% of those from reliable sources, and reactions were about 3 times greater. The relative amplification coefficient, the ratio between the mean reactions per post from questionable sources and from reliable sources, is about 1 on Twitter, 0.35 on YouTube, 0.55 on Reddit, and 3.9 on Gab. The authors conclude that information from reliable and questionable sources does not differ significantly in its diffusion patterns, and that Gab is

the environment most susceptible to misinformation. Furthermore, what drives diffusion are the peculiarities of each platform and the group dynamics of those who engage with the topic. We read this as a sign of how strong the effect anchored in the history of those who pass the information on can be; the study itself does not measure that history.

The limits are those of the data collection. The study covers the first 45 days, and the first Twitter data point is from January 27, because of the limits of the platform's application programming interface; Twitter's streaming collection represents about 1% of the volume, and Instagram's was done by hand. Amplification is relative to each platform and depends on the outlet label, not on the veracity of each claim. And the authors note, in the introduction, that studies had pointed out that false information may spread faster and wider, but that this might be a platform-specific effect, which their data do not contradict.

For Brazil, Massarani et al. (2021) used BuzzSumo, a monitoring tool that shows which content on a topic generated the most engagement on social media, to collect the hundred links (the maximum the tool allows) containing the word "vacina" (vaccine) that had the highest engagement in 2020. Of these, 94 valid links in Portuguese (after excluding five links from Portuguese outlets and one broken link) were compared with those from May 22, 2018, to May 21, 2019, which yielded 89 links, with the same filters. The authors operationalized engagement as the sum of shares, comments, likes, and reactions on Facebook, shares on Twitter, shares and comments on Reddit, and shares on Pinterest, but noted that the definition was restricted to the networks' protocols. They classified the accuracy of each link, comparing the information cited with scientific articles, official sources, and other news outlets, as verified, unverifiable (such as personal opinions, polls, and fictional content), or disinformation, the last subdivided into types such as misleading content, fabricated content, and false connection, with categories adapted from a proposal by other authors. They also classified the source of each link: professional, if it had a defined editorial policy, a responsible institution, identified authors and staff, and citation of sources, or nonprofessional. Table 11.3 brings the numbers together.

Mean engagement per link increased 8.6 times, from 41,405 to 354,321. The proportion of disinformation practically did not change, nor did that of verified information. What changed was the profile. In 2018–2019, disinformation was mainly fabricated content, that is, 100% false and created to deceive or harm (66.7%), issued by nonprofessional outlets (83.3%); in 2020, most of it was false connection (84.6%), and professional outlets became the main

sources (61.5%). The mean engagement of disinformation, which in 2018–2019 was lower than that of verified content (36,200 versus 40,483), was higher in 2020 (412,223 versus 338,019).

The authors' reading is that the greater engagement of disinformation suggests a greater potential for going viral during the pandemic, and they link false connection to sensationalist headlines and to dependence on engagement-based advertising revenue, drawing on other works. It is recorded as their interpretation: the article measures neither revenue nor the editing of headlines. The reading converges with the hypothesis of Chapter 8, that feedback in the form of attention selects headline forms, and this study does not test it. From the percentages, 2018–2019 corresponds to about 12 items of disinformation and 2020 to about 13, a calculation reconstructed here, but the article relies on descriptive measures. Differences in means between 12 and 13 items, with the variability of engagement that social media show, do not support much more than the description presented.

TABLE 11.3

Vaccine links with the highest engagement on social media, 2018–2019 and 2020

Measure	2018-2019	2020
Valid links	89	94
Total engagement	3,685,100	33,306,200
Mean engagement per link	41,405	354,321
Verified information	80.9%	81.9%
Disinformation	13.5%	13.8%
Unverifiable	5.6%	4.3%
Mean engagement: verified	40,483	338,019
Mean engagement: disinformation	36,200	412,223
Type of disinformation: fabricated / false connection	66.7% / 8.3%	15.4% / 84.6%
Main source of disinformation	Nonprofessional outlets (83.3%)	Professional outlets (61.5%)

Note. Values from the article. In 2018-2019 there was also 25% misleading content among the disinformation; in 2020, none. "False connection": titles, headlines, and visual elements that contradict the content (Box 1 of the article).

The authors see, in both periods, political instrumentalization of the debate on vaccines, with examples from different camps. In 2018–2019 they describe false content and fact-checks about the meningitis vaccine and a false item attributing to a foreign politician a statement he did not make; in 2020, CoronaVac was the vaccine most cited in disinformation content (23.1%), although it did not lead overall mentions, and the authors read the doubts about it as grounded in political issues rather than scientific evidence. The reading is the

authors': the article measures which vaccines were cited, not why those who cite them do so. A return to Chapter 1 (Section 1.6) shows that it brings together the other Brazilian data on disinformation.

DATA	In the first 45 days of 2020, the growth pace of posts and reactions from questionable and reliable sources was the same within each platform, and relative amplification ranged from about 0.35 to 3.9 (Cinelli et al., 2020). Among the vaccine links with the highest engagement in Brazil, mean engagement increased 8.6 times between 2018–2019 and 2020, the proportion of disinformation remained practically the same, and its profile changed, from content fabricated by nonprofessional outlets to headlines that contradict the content, issued mainly by professional outlets (Massarani et al., 2021).
INTERPRETATION	In both studies, the COVID-19 infodemic appears as an increase in volume and engagement and a change in profile, not as an increase in the proportion of falsehood. The platform and the period change the result.
HYPOTHESIS	The change in profile, from fabricated content by nonprofessionals to misleading headlines by professionals, reflects the contingency of feedback in the form of attention to which professional outlets are subject (Chapter 8). This is the interpretation of the study's authors, without a test; it would be necessary to measure the editing of headlines before and after.

11.6 What infodemiology proposes to do

Eysenbach (2020) proposes four pillars of infodemic management: information monitoring (infoveillance), building health and science literacy, refining knowledge through fact-checking, peer review, and similar processes, and translating knowledge in a timely manner and without political or commercial distortion. He describes the crisis at four levels: science, policy and practice, news media, and social media, in layers sized according to how much information each group of actors produces, and he says that the model is expert opinion, not the WHO framework.

The WHO consultation reported in the article by Calleja et al. (2021) brought together 110 participants from over 35 countries and produced, through structured expert judgment, a research agenda with five workstreams, 166 suggested questions, and 49 high-priority ones. The workstreams are measuring and monitoring impact; detecting signals and understanding spread and risk; responding with interventions; evaluating interventions and strengthening resi-

lience; and promoting the development and application of interventions and toolkits. The authors say that the current approaches to communication, community engagement, and messaging are not enough, on their own, to manage the harmful effects of the infodemic, and that the response requires intervention at multiple levels: individual, community, medium, platform, and policy. They write that communication campaigns can raise the visibility of messages but are often not effective at debunking false claims, and that half of fact-checking organizations do not work with health professionals. Among the challenges of the information environment, they list the computational amplification of polarizing messages, microtargeting, and the weakening of local journalism.

Zielinski (2021) adds two positions. The first is that the distinction between *misinformation* and *disinformation* should be kept, since the solutions are likely to differ: large doses of good information help against the former, while the latter uses the same channels as good information. He also distinguishes 'malinformation,' information grounded in reality but used to cause harm, a Council of Europe definition that the article reproduces. The second is that remedial actions need human rights safeguards, so that they do not end in censorship and restrictions on freedom of expression, and that technology alone will not provide a solution.

The chapter carries three points into Part IV. The first is the distinction among the things an intervention can alter, and the definition of reach, exposure, and reaction organizes the question: the platform's delivery rule (Chapter 12, Section 12.6) alters reach; signals and training alter reaction (Sections 12.2 to 12.5); taken together, acting on the consequences of the reach of and exposure to disinformation can reduce harm, and reaction can feed back into delivery itself. The second is that the field's agenda is one of questions, not answers: the WHO agenda is the product of expert judgment, and here it is read as a map of what is unknown; and what is unknown may involve conditions that would steer responses toward a lower degree of disinformation. The third is that none of the diffusion studies read measures what the information did to people's health behavior, such as getting vaccinated or adhering to a measure. They measure diffusion, and diffusion is the outcome the field measures most easily. This limits what can be said about the effectiveness of any intervention evaluated only by reduced diffusion, and Chapter 12 returns to this limitation when evaluating interventions.

11.7 Questions for reading a diffusion study

For those who use the chapter to read studies, news reports about studies, or platform reports, the questions that the preceding sections suggest are the following.

1. What is the unit being compared: fact-checked claim, link, outlet, topic, account?
2. Who classified the true and the false, and by what criterion (fact-checker, outlet label, the authors themselves)?
3. What is being counted: reach, exposure, or reaction? What was inferred, and from what?
4. Is the measure one of level (how much) or of structure (how, over how many generations)? Final size says little about structure.
5. What are the basis of comparison, the platform, the language, and the period?
6. Does the study measure what the information did to the behavior of those who saw it, or only how far it spread?

Summary

Infodemiology is, in Eysenbach's proposal, the science of how information is distributed in an electronic medium and what determines that distribution, with the aim of informing public health; an infodemic, in WHO usage, is an overflow of information of varying quality during an acute health event. For Zielinski, an infodemic is defined by volume and velocity, and falsehood is a secondary characteristic. The Lombardy case, as reported by the authors who tell it, shows that information that is not false can have harmful effects because of its timing and what people do with it (11.1).

Measures of diffusion count mainly reaction (reposting, retweeting, engaging), infer exposure from the list of followed accounts, and leave reach to the platform's rule. Structural virality, the average distance between the pairs in a cascade, shows that, on Twitter in 2011–2012, almost everything is small and almost everything is broadcast, and that the size of a cascade says little about its structure (11.2). Falsehood went farther and faster than truth on Twitter from 2006 to 2017, among fact-checked claims, and two other studies, with other units, do not yield the same result; the reviews the plan anticipated were not read (11.3).

Epidemiological models describe growth and make it possible to rule out hypotheses, such as that of the tipping point, but they treat the resharing response as an event triggered by exposure, with no term for the consequence that maintains it; scale belongs to the level of the population, and the explanation, in the reading proposed here, lies in contingencies between people (11.4). In COVID-19, in the first 45 days of 2020, the growth pace was the same for reliable and questionable sources within each platform, and among vaccine links in Brazil engagement increased 8.6 times, with the proportion of disinformation practically unchanged and its profile altered (11.5). The field proposes to monitor, build literacy, refine, and translate, at multiple levels, and its agenda is one of questions; no diffusion study read measures the effect on what people do (11.6).

The distinction among reach, exposure, and reaction, which Chapters 8 and 10 announced, is fixed in Section 11.2 and serves to read the measures of Chapters 8 and 9, which count one of the three without separating them, and to support Chapter 10's caveat that a post's effective window depends on who sees it. Chapter 12 starts from what was left open here: what can be changed, at four levels, and with what evidence.

References

Calleja, N., AbdAllah, A., Abad, N., . . . Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference. *JMIR Infodemiology*, *1*(1), Article e30979. https://doi.org/10.2196/30979

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), Article 16598. https://doi.org/10.1038/s41598-020-73510-5

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, *11*(1), Article e11. https://doi.org/10.2196/jmir.1157

Eysenbach, G. (2020). How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research*, *22*(6), Article e21820. https://doi.org/10.2196/21820

Glenn, S. S. (2004). Individual behavior, culture, and social change. *The Behavior Analyst*, *27*(2), 133–151. https://doi.org/10.1007/bf03393175

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), Article e2025334119. https://doi.org/10.1073/pnas.2025334119

Lazer, D. M. J., Baum, M. A., Benkler, Y., … & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Massarani, L. M., Leal, T., Waltz, I., & Medeiros, A. (2021). Infodemia, desinformação e vacinas: a circulação de conteúdos em redes sociais antes e depois da COVID-19 [Infodemic, disinformation, and vaccines: The circulation of content on social media before and after COVID-19]. *Liinc em Revista*, *17*(1), Article e5689. https://doi.org/10.18617/liinc.v17i1.5689

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), Article 7144. https://doi.org/10.1038/s41467-022-34769-6

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zielinski, C. (2021). Infodemics and infodemiology: A short history, a long future. *Revista Panamericana de Salud Pública*, *45*, 1. https://doi.org/10.26633/rpsp.2021.40



PART IV

Interventions

12 Interventions: What Can Be Changed, at Four Levels, and with What Evidence

CHAPTER

12

Interventions: What Can Be Changed, at Four Levels, and with What Evidence

A field study with YouTube ads serves as a starting point. An inoculation video, watched for 30 seconds or more, was followed, within 24 hours, by a single question: which manipulation technique a fictitious post used. The proportion of correct answers was higher among those who saw the ad than in a comparison group, and the authors describe the result as an average increase of about 5% in recognition. The effect size, measured by Cohen's h (a measure of the difference between two proportions, in which 0.2 is, by convention, a small effect), was 0.09, below the smallest effect size of interest that the authors themselves had preregistered (0.10) (Roozenbeek et al., 2022).

It is one of the most careful intervention studies among those read: six randomized experiments, the first five with preregistered hypotheses, and a field study with more than 22,000 responses. That is also why it is a good starting point. What it measures is the recognition of a technique by people who saw an ad. It does not measure whether a person shares less, nor what that person's audience does with what he or she shares, nor what happens a week later; the authors state that the field study assessed susceptibility, not behavior. Between what the study measures and what Chapters 3 to 11 treated as what maintains false speech, or speech classified by an outside party as hate speech, there is a distance, and this chapter exists to measure it, intervention by intervention.

The analysis in the previous chapters located the maintenance of speech in contingencies, not in information: the audience that reacts (Chapter 9), the delivery rule that decides reach (Chapter 8), the verbal community that reinforces and protects the speaker (Chapter 6), and the cultural practices that

bring all of this together (Chapters 9 and 10). Each suggests a point of intervention, and the chapter organizes them into four levels, close to those listed in the WHO research agenda (individual, community, media, platform, and policy; Calleja et al., 2021): the individual, which is what a person reads, sees, or learns; interaction and the group, which is what is done in the face of the speaker; the platform, which is what the delivery rule and moderation change; and culture, which is practices and metacontingencies.

Figure 12.1 sets side by side eleven intervention studies read in full and states, for each, what was measured. The count is our own, made from the figure. Eight studies measured judgment or task performance; five, behavior observed on a platform or forum; four, observed behavior with a follow-up of one week or more; only one, Guess et al. (2020), measured all four; and none evaluated an intervention at the cultural level. The distribution is not accidental. The individual is the level at which testing is cheap, because one shows someone some material and asks what they thought of it. The other levels require access to accounts, forums, and ranking, and it is at those levels that the measure becomes what people do.

Three cautions guide the reading. The first is that no intervention was tested here. The studies read are by others, interpreted in light of operant analysis, and when it is said that an intervention "teaches discrimination" or "acts on the contingency," the sentence is a reading proposed here and is marked as a hypothesis. Nor were the later reviews and replications on correcting claims read, nor the Twitter study that one of the cited articles describes (Section 12.3); where this matters, the text says so. The second is measurement. Chapter 1 showed that a survey response is behavior under the control of the situation in which it was given, and Chapter 11 that no diffusion study read measures what information did to people's behavior. The same holds here: judging a headline on a screen is not sharing it, and sharing it is not changing what one does; the correspondence between saying and doing, which Chapter 7 (Section 7.4) treated as an effect of contingencies, not as a given, applies to the stated intention to share. The third is the criterion. Chapter 10 used the same one for what widens and for what narrows the Overton window, and this chapter uses it for interventions: the same question, the same kind of measure, and the same caveat, whatever the political camp of whoever emits the content one wants to reduce.

The path follows the levels and, within the individual, the timing of the intervention. Section 12.1 deals with correcting afterward, through correction and warnings; 12.2, with correcting beforehand, through inoculation read as discrimination training; 12.3, with stimuli that signal accuracy. Section 12.4 deals with what is done in the face of the speaker (counterspeech, public commit-

ment, and the hypothesis of source autoclitics); 12.5, with reducing prejudice through equivalence and through comics; 12.6, with platform contingencies. Section 12.7 brings together what is meant here by cultural design and the ethical limits of those who design.

FIGURE 12.1

What each intervention study measured

	Judgment or task performance	Stated intention to share	Observed behavior (platform or forum)	Follow-up of one week or more
Individual: what the person reads, sees, or learns				
Nyhan and Reifler (2010) *correction within a mock news article*	●	–	–	–
Freeze et al. (2021) *retrospective misinformation warning*	●	–	–	–
Roozenbeek and van der Linden (2019) *inoculation game*	●	–	–	–
Roozenbeek et al. (2022) *inoculation videos*	●	●	–	–
Guess et al. (2020) *tips for spotting false news*	●	●	●	●
Dixon and Lemke (2007) *equivalence training (adults)*	●	–	–	–
Mizael et al. (2021b) *equivalence training (children)*	●	–	–	●
Interaction and group: what is done in the face of the speaker				
Hangartner et al. (2021) *empathy-based counterspeech, on Twitter*	–	–	●	●
Tsipursky et al. (2018) *public commitment to the truth*	–	–	●	●
Platform: what the delivery rule and moderation change				
Álvarez-Benjumea and Winter (2018) *removal of hostile comments (simulated forum)*	–	–	●	–
Nyhan et al. (2023) *demoting like-minded sources (Facebook)*	●	–	●	●
Culture: practices and metacontingencies				
No culture-level intervention was evaluated in the studies read for this book.				

● *the variable was measured (the result may have been null).* – *not measured.*
Original map, based on the studies read (Chapters 4 and 12). Follow-up: Guess et al., about 20 days (U.S.); Hangartner et al. and Tsipursky et al., four weeks; Nyhan et al., three months; Mizael et al., six weeks. Measurement on the same day or within 24 hours does not count as follow-up.

Note. Original map, based on the studies read (Chapters 4 and 12); it presents no new data. A mark records that the variable was measured, and the result may have been null. Follow-up: about 20 days in Guess et al. (in the United States), four weeks in Hangartner et al. and in Tsipursky et al., three months in Nyhan et al. (2023), and six weeks in Mizael et al. (2021b); measurement on the same day or within 24 hours does not count as follow-up. The reading of each cell is in sections 12.1 to 12.6.

12.1 Correcting afterward: correction, warnings, and the "backfire effect"

The most common intervention in the face of a false claim is correction: after the claim has circulated, someone refutes it. Nyhan and Reifler (2010) tested it in four online experiments with undergraduate students at a Midwestern

Catholic university in the United States, in fall 2005 and spring 2006. Participants read a mock news article that included a misleading claim from a politician, with or without a correction, and then said whether they agreed with it. The authors chose a between-subjects design to maximize the effect of the corrections, because measuring before and after would anchor the answers to the first one. They define misperceptions as false or unsubstantiated beliefs about the world, and they call it backfire when a person comes to support the original opinion even more strongly.

The results are not uniform. In Study 1 (130 participants), the correction from the Duelfer report on weapons of mass destruction in Iraq did not reduce misperceptions overall; among conservatives (n = 33), agreement that Iraq had such weapons rose from 32% in the control condition to 64% with the correction, and among non-conservatives (n = 97) it went from 22% to 13%. In the tax-cut experiment, conservatives' agreement that the cuts increased revenue went from 36% to 67% with the correction (n = 60), and that of non-conservatives from 31% to 28% (n = 136). In two other experiments, the result was different: in the second version of the weapons of mass destruction experiment, the correction left conservatives more inclined to believe that Iraq lacked those weapons, and the authors say it is unclear why; in the stem-cell experiment, the correction worked for conservatives and moderates, but not for liberals, without backfire. Attributing the article to the *New York Times* or to FoxNews.com left the correction's effect unchanged in all three experiments of Study 2.

The authors place the caveats in the article itself. All of the backfire results came from conservatives, and without more studies it is impossible to know whether liberals and conservatives react differently to corrections; it would be valuable to replicate with a non-college sample. A post hoc analysis, which the authors describe as pushing the data to the limit, suggests that the correction failed among conservatives for whom Iraq was the most important issue, but only 34 respondents (eight right of center) named it. And they write that present-day political misperceptions are a moving target, liable to shift rapidly, as suggested by the difference between the two weapons of mass destruction experiments.

Correction also arrives as a warning. Freeze et al. (2021) studied what they call the "tainted truth" effect: a misinformation warning applied to authentic news, which reduces its credibility and impairs political memory. In an online experiment with 434 MTurk participants, all watched a four-minute C-SPAN video of floor speeches by members of Congress, answered 22 questions unrelated to it, and were randomly assigned to six conditions. They read a fabricated news article about the video, either vague (control), with the correct

facts, or with some of the facts altered, and either did or did not receive the warning "some of the information presented in the news article you read was inaccurate." The authors call the warning valid when it correctly targets misinformation, and invalid when it falls on an article with no altered facts. Afterward, everyone took a recognition test of 20 facts from the video and rated the credibility of the video and the article. The warning made the article seem 0.59 points less credible (on a 5-point scale), in all conditions, including when the article was accurate. The effect on memory was significant in only one of the three tests.

DATA	In four experiments with students, the correction embedded in the news article made conservatives' agreement with the misleading claim rise in two (weapons of mass destruction, Study 1; tax cuts) and worked among them in the other two (weapons, second version; stem cells, for conservatives and moderates but not for liberals). Among non-conservatives, agreement fell on the two topics for which the figures are reported (Nyhan & Reifler, 2010).
INTERPRETATION	The study measures agreement with a claim, in a survey, right after reading. It does not say what the person does with the claim or how long the response lasts. The backfire results come only from misperceptions held by conservatives, on the topics chosen, and the authors say they cannot claim that the political camps differ; the study supports neither "conservatives reject corrections" nor "corrections backfire."
HYPOTHESIS	The authors read backfire as counterarguing information that is incongruent with one's preferences (their interpretation, untested). Two rival readings are added here, equally untested in this design. The first is that of Chapter 6: the correction is an aversive stimulus for those whose claim is tied to their community, and disagreeing more is a countercontrol response, or avoidance of it. The second is that of repetition (Chapter 3, Section 3.5): the correction re-presents the claim, and instructions such as "don't believe that all members of group X are Y" may increase the salience of the group and add attributes to the category (Dixon et al., 2003). Separating the three readings would require varying whether the person had previously stated the position in public and whether the correction repeats the statement or not.

The valid retrospective warning, given after reading, helped little. The mean memory score was 59% in the control condition, 46% with misinformation and no warning, and 52% with misinformation and a valid warning, and the

misinformation score went from 33% to 25%; the difference between 46% and 52% was not significant at 0.05 ($p = 0.06$). The authors conclude that, in light of their findings, broad misinformation warnings are best avoided, since using them across the board can lower the credibility of legitimate news sources and lead people to throw away useful information. They also note that the warning came from the researcher, with no clearly specified source, and that it would be interesting to see whether motivated reasoning alters the effect when it comes from another source.

Read here, the result of Freeze et al. has the form of stimulus generalization. A warning without a target ("this may be misinformation") does not say which relation in the article is at stake, and the reader responds by reducing confidence in the whole, including what is valid. A specific warning, one that located the statement and the reason, would set the occasion for a different response. This is a hypothesis, and the study does not test it: all the warnings were general, and none varied specificity. The measure, once again, is one of judgment: credibility and memory, right after reading.

What the section leaves as data is little, and it is useful for that reason. Correcting afterward, in the design of the studies read, has an uncertain effect, which varied with the topic and did not vary with the source attributed to the article in the experiments in which this was manipulated, and was not measured beyond the session. The hypothesis that correction is aversive for those whose claim is tied to their community (Chapter 6) is compatible with the results, and no study has tested it. The next section deals with what has been done to prevent, rather than correct.

12.2 Correcting beforehand: inoculation as discrimination training

Inoculation reverses the order of correction. Instead of refuting a claim after it has circulated, the person is exposed, beforehand, to a weakened version of the attack. Roozenbeek et al. (2022) define psychological inoculation by two components: a forewarning, which leads people to perceive a threat that their attitudes are about to be attacked, and exposure to a weakened microdose of misinformation containing a preemptive refutation. The authors attribute the theory to McGuire, in the 1960s, and it follows a medical immunization analogy; it is not a behavior-analytic concept, and the reading the chapter makes of it is proposed here.

The first study read is a game. In *Bad News*, the player acts as a producer of fake news and learns six techniques for producing misinformation: polarization, invoking emotions, conspiracy theories, trolling, deflecting blame (dis-

crediting opponents, which deflects attention away from accusations of bias by attacking the source of the criticism), and impersonating accounts. The game is free, runs in a browser, and takes about 15 minutes; using the strategies earns the player followers and credibility, and choosing options consistent with ethical journalistic conduct costs credibility or followers (Roozenbeek & van der Linden, 2019). The sample was a convenience sample, recruited through a university press release, with 14,266 complete pre-post pairs collected over three months in 2018; it was self-selected and skewed toward men (75%) and toward somewhat-to-very-liberal individuals (59%).

The measure was the reliability, from 1 to 7, of tweets and headlines rated before and after the game, in six items, two of them control items of real news. The reliability attributed fell for impersonation (from 3.00 to 2.30; d = 0.36), conspiracy (from 2.47 to 1.97; d = 0.35), and discrediting (from 2.37 to 1.92; d = 0.30), and the control items barely changed, which the authors read as an indication that the game does not merely make participants more skeptical. The polarization item, collected post hoc because of a technical storage failure (885 pairs), had a smaller effect (d = 0.16). The average effect on the false items was from 2.61 to 2.06 (d = 0.52) and was larger among those who rated them as more reliable at pretest.

The authors flag the limitations. For ethical reasons they did not randomize who would play: there is no traditional control group, and they say that alternative explanations cannot be completely ruled out; future studies should model the decay of the effect, to find out whether and when booster doses would be needed. One caveat should be added: the larger effect among those who rated the items as more reliable at pretest is also the pattern expected from regression to the mean, and without a control group there is no way to separate the two.

The 2022 study is of another scale. Five short videos deal with emotionally manipulative language, incoherence, false dichotomies, scapegoating, and *ad hominem* attacks; each contains a forewarning, a preemptive refutation of the technique, and a microdose in innocuous, humorous, nonpolitical, and fictitious examples. In studies 1 to 6, participants were randomly allocated to either a 1.5-minute inoculation video or a neutral video, and then rated ten synthetic posts, manipulative or neutral, on four measures, among them willingness to share. The main outcome is discrimination (*discernment*, in the authors' term): the difference between the averaged scores for the manipulative and the neutral stimuli. The manipulative stimuli were not always false; they were built to use a manipulation technique in an obvious way.

In studies 1 to 5 (n = 5,416, national quota samples of the United States), the preregistered hypotheses were supported for 16 of 20 outcome measures. Cohen's d for technique recognition ranged from 0.28 (scapegoating) to 0.68 (false dichotomies) and, for sharing, in the form of discrimination, from 0.10 (incoherence, not significant) to 0.22 (false dichotomies). A year later, study 6 (n = 1,068) replicated the emotional language video, with d = 0.34 for sharing. For the smaller effect in the scapegoating and incoherence videos, the authors' explanation is that the control group already distinguished these techniques better, that is, the baseline was higher.

In the field study, two videos ran as ads on YouTube, which viewers could skip; about 967,000 users watched one of them for 30 seconds or more, and a random 30% received, within 24 hours, a single test question within YouTube itself, in which they had to identify the technique used in a fictitious post (four options, only one correct). The comparison group, matched on demographics, did not see the video and answered one of the questions; 22,632 responses were collected. The proportion of correct answers was higher in the treatment condition (h = 0.09, below the preregistered smallest effect size of interest, h = 0.10). The authors report an average cost of about US$0.05 per view and say that about 5.4 million people saw the videos over 15 days.

The stated limitations are the ones that matter for the reading. The effect was measured within 24 hours (median of 18.4 hours), its duration was not studied, the videos and the samples were from the United States, and the field study assessed susceptibility, not behavior. Each field participant saw a single test item, and the pairs of manipulative and neutral stimuli are not exact mirrors of each other and may differ on other dimensions. The authors summarize, citing a study that was not read, that prompting attention to accuracy led Twitter users to share more high-quality sources, and they say that how to reduce susceptibility in real-world settings remains an open question. They found no consistent interactions between condition and the demographic and ideological covariates.

The reading proposed here is that inoculation, in the 2022 study, has the form of discrimination training with exemplars of the technique, not of the content. The property common to the examples is the technique (emotional language, the false dichotomy); the content varies and, in the test, is different. What is measured is the difference between the rating of a post that uses the technique and that of a neutral post with the same content, and this is what the concept of discrimination calls for. Figure 12.2 puts the procedure in this form, with what the study measures in solid lines and what is added here in dashed lines.

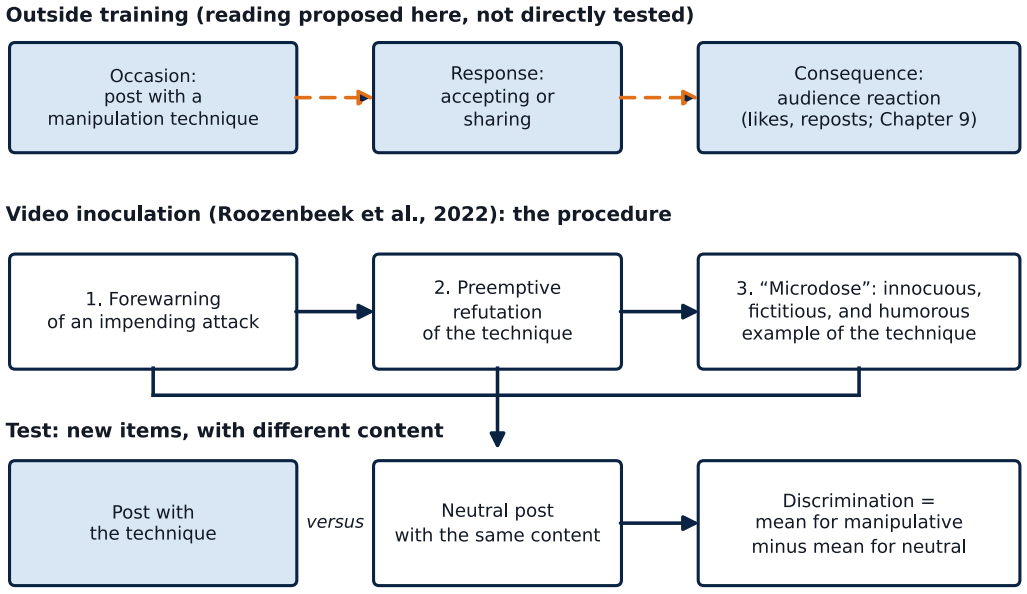
CONCEPT IN USE

Discrimination training (Skinner, 1957; Shea et al., 2023)

In all verbal behavior under stimulus control, the prior stimulus marks the occasion on which the response tends to be reinforced and, through operant discrimination, comes to be the occasion on which the response tends to be emitted (Skinner, 1957). In the tact, control is established by reinforcing the response, as consistently as can be managed, whenever a single stimulus is present, using many different reinforcers or a generalized reinforcer, and the control that results is exerted by the stimulus. *Multiple exemplar training*, as described by Shea et al. (2023), presents a set of stimuli that samples the range of target stimuli and response topographies, and continues until the response comes under the control of the features of the examples and also of novel stimuli that share them. Reading proposed here: discrimination training is the procedure that brings a response closer to a property of the stimulus, not to the particular stimulus; the test that it worked is the response to new items that share the property, with different content.

FIGURE 12.2

Video inoculation read as discrimination training



Original diagram, without data. Solid line: the study's procedure. Dashed line: reading proposed here, which the study's measure does not reach (the field study did not measure sharing; the social consequence is not in the training).

Note. Original diagram, without data. Solid lines: the study's procedure. Dashed lines: reading proposed here, which the study's measure does not reach; the field study did not measure sharing, and the social consequence is not part of the training.

The reading has two weak points, which should be stated. The first is that, in the videos, there is no differential consequence programmed for discriminating: the viewer watches, listens, and responds afterward. What the procedure has is instruction and example, and the term "training" is, to a large extent, an analogy. In *Bad News* there is a consequence, in followers and credibility, but it is for using the technique, not for recognizing it; the test measures the other repertoire, that of rating, and transfer from one to the other is an assumption. The second is that the dashed band of Figure 12.2 is the one that matters for the analysis: outside the study, the technique appears on an occasion when there is an audience, and the response of sharing has the consequence that Chapter 9 described. Nothing in the training touches that consequence.

DATA	In the game *Bad News*, without a control group, the reliability attributed to items with three of the techniques fell with d from 0.30 to 0.36, and that of the real items did not change (Roozenbeek & van der Linden, 2019). In the videos, six randomized experiments supported 16 of 20 preregistered hypotheses in studies 1 to 5, and the field study had h = 0.09, below the smallest effect size of interest, measured within 24 hours, in the United States, and as susceptibility (Roozenbeek et al., 2022).
INTERPRETATION	Brief exposure to examples of a technique improves, shortly afterward, the rating of new items that use it, without merely making the person more skeptical. The effect is small in the field, and what is measured is judgment, not behavior.
HYPOTHESIS	Inoculation works as multiple exemplar training with the technique, and the rating response generalizes to new content (reading proposed here). A rival reading is that the viewer received a rule and responds to it while the rule is salient. Both predict the effect shortly afterward; they differ in duration and in what happens when the rule competes with the audience's reaction. Separating them would require measuring, with follow-up, sharing on a platform where the same post does or does not receive a reaction.

12.3 Stimuli that signal accuracy

A third family of interventions, besides correction (Section 12.1) and inoculation (12.2), neither teaches beforehand nor corrects afterward: at the moment the person reads, it presents a stimulus that signals what to look at. The one read in detail is the list of tips that Facebook promoted at the top of the news feed in 14 countries in April 2017, developed with the nonprofit First Draft: "Tips to Spot False News," with 10 strategies in the U.S. version and 6 in the

Indian one. Guess et al. (2020) describe it as a set of simple decision rules for telling mainstream news apart from false news, an ability they call discernment (here, discrimination), without requiring readers to conduct research or think deeply about each piece of news.

In the United States, the study was a two-wave online panel, conducted by YouGov after the 2018 midterm elections (wave 1, N = 4,907; wave 2, N = 4,283), with participants randomly assigned to read or not read the tips. The main measure was the perceived accuracy (from 1 to 4) of mainstream, hyperpartisan, and false headlines that had circulated in real outlets or on Facebook or WhatsApp in the previous six months, some of them rated as false by at least one fact-checking agency. In the first wave, the tips reduced the perceived accuracy of false news by nearly 0.2 points on the 4-point scale (ITT = −0.196, the effect on everyone assigned) and by nearly 0.3 among those actually treated (ATT = −0.299); about 66% passed the comprehension questions.

The intervention slightly lowered belief in mainstream news (ITT = −0.046), but it increased discrimination between mainstream and false news (ITT = 0.146), a 26.5% improvement relative to the control group. With a median of 20 days between waves, the effect on the perceived accuracy of false news remained distinguishable from zero, but attenuated by more than half (ITT = −0.080). On the binary measure, the proportion who rated false headlines as very or somewhat accurate fell from 32% in the control group to 24% in the first wave and, in the second, from 33% to 29%.

In the online sample from India, the intervention reduced the perceived accuracy of false news (ITT = −0.126) and improved discrimination by 17.3%; only 28% passed the comprehension questions, versus 66% in the United States. In the face-to-face sample, drawn at random from the voter registration list of four constituencies in Uttar Pradesh and administered orally, the authors found no evidence of an effect; 11% of that sample used WhatsApp, versus 90% of the online sample. In the second wave, they found no reliable evidence of an effect in either of the two Indian samples.

What the study did not find matters as much as what it found. The authors say the evidence is insufficient to show that the intervention changed either the stated intention to share false news or the behavioral measures of news consumption; intentions to share mainstream news increased, and those to share hyperpartisan news decreased. Among the caveats: the effects were modest and decayed; mainstream news came to be rated as slightly less accurate; many real Facebook users probably ignored the link; and the lack of change in

actual consumption may be because information habits are ingrained. Using web-tracking data, they estimate, in a calculation they describe as hypothetical, that false positives would fall from 6.1% to 4.9% of the stories consumed.

Two comments by the authors situate the intervention. They compare the effect to responses to Facebook's "disputed" labels in an earlier study (ITT = −0.236) and to a general warning (ITT = −0.079); those studies were not read. And they note that trying to identify false news through close scrutiny of the headline differs from the approach of professional fact-checkers, who use lateral reading of other sources, and that future studies should clarify whether the effect comes from increased accuracy concerns or from more effective heuristics.

There are two other accuracy-signaling results in the sources read. In one of the studies by Ceylan et al. (2023), judging the accuracy of a headline before deciding whether to share it reduced the sharing of false headlines, but it did not increase discrimination among the more habitual users more than among the less habitual ones; the study was conducted with 839 participants, and the task was a simulated share button. And Freeze et al. (2021), in Section 12.1, show the price of a signal without a target.

DATA	In the United States, the list of tips reduced the perceived accuracy of false news by about 0.2 points (out of 4), increased discrimination by 26.5%, and the effect fell by more than half after a median of 20 days; in India, the effect was smaller in the online sample and was not found in the face-to-face sample or in the second wave. There was insufficient evidence of change in the stated intention to share false news or in observed consumption (Guess et al., 2020).
INTERPRETATION	The tip changes the judgment of a headline, for a short time, where it was understood (66% in the United States, 28% in the Indian online sample) and where the channel of circulation is the study's (11% WhatsApp use in the face-to-face sample, 90% in the online one). It does not change, in the data, what people do with the news.

HYPOTHESIS The tip is a verbal stimulus that describes a contingency: a rule, in the vocabulary of Chapter 3 (Section 3.7). Rules control while they are salient and the response is cheap; with no consequence programmed for applying them, and with the audience's reaction (Chapter 9) maintaining the response of sharing, the decay the authors found is to be expected, and habit, in Ceylan et al.'s sense, leaves little room for them. This is a reading proposed here; the mechanism, the authors themselves say, remains to be clarified.

12.4 In the face of the speaker: counterspeech, public commitment, and source autoclitics

Between the reader and the platform that delivers lies interaction: what someone does, on the same network, in the face of a person who emitted a false statement or one classified, by an outside party, as hate speech. Two studies read evaluate interventions of this kind. The section ends with a hypothesis proposed here that no study tests.

Hangartner et al. (2021) conducted a field experiment on Twitter. Using a dictionary of xenophobic terms, sentiment analysis, and manual classification, they selected users who had sent xenophobic or racist tweets in English, excluding verified accounts, retweets, possibly sarcastic tweets, and accounts that were probably bots or minors. Each user was randomly assigned to one of three one-shot counterspeech messages (empathy, warning of consequences, humor) or to the control arm (40%), with no intervention; the counterspeech came, within 24 hours, from a neutral, human-controlled profile, created about four weeks earlier, with an apolitical posting history and about 100 followers. The empathy message sought to elicit empathy (saying, for example, that it is painful to see this kind of language); the warning message reminded the user that friends and family can see the tweet; the humor message used an animal meme with a caption asking the user to stop tweeting. The fieldwork ran from November 23, 2020, to January 17, 2021, and the analytic sample, after the loss of 230 accounts that were set to private or suspended, was 1,350 users (humor 284, warning 269, empathy 259, control 538).

The analysis was by intention to treat, that is, each user counts in the group to which he or she was assigned, whether or not the message was received (5.5% of the tweets were deleted by their senders between assignment and treatment), with separate regressions for each of the six outcomes. Because there are many outcomes and treatments, the authors also report adjusted P values (by the Benjamini–Hochberg method, which corrects for the excess of chance significant results that multiple tests produce). Empathy reduced the number

of xenophobic tweets over the following four weeks by 0.10 standard deviations (P = 0.056; adjusted P = 0.112) and the total number of tweets by 0.13 (P = 0.039; adjusted = 0.112), and it increased deletion of the original tweet by 0.21 (P = 0.009; adjusted = 0.053), about 8.4 percentage points. In counts, the average user in the empathy group sent 1.3 fewer xenophobic tweets and, in total, 91.6 fewer tweets. The joint F test, which examines all outcomes at once, rejected the null hypothesis of no effect for empathy (P = 0.017), but not for humor (P = 0.285) or for warning of consequences (P = 0.633).

The authors say that, contrary to the preregistered hypotheses, they found no significant empathy effects on the other three outcomes, and that the separate analysis of the two components of the xenophobic tweet share is exploratory; the preregistered outcome, the share itself, did not change. They conclude that the results temper expectations about the most common counterspeech interventions used by organizations, and they recommend studying repeated interventions and paying attention to *chilling* effects "beyond hate speech," in their words. The study was approved by an ethics committee and preregistered. The premise, which the authors support with other work, is that content moderation can silence legitimate speech along with illegal speech and scatter, instead of diminishing, what they call hate speech, and that counterspeech seeks to reduce hate by persuading whoever emits it, rather than to suppress expression.

The second study is of a different nature. Tsipursky et al. (2018) describe the *Pro-Truth Pledge*, a proposal in which signers, private citizens and public figures, commit to 12 behaviors correlated with an orientation toward truthfulness, grouped into sharing, honoring, and encouraging truth; among them, verifying information before accepting and sharing it, retracting it if one cannot verify it, and asking allies to retract refuted information. In the authors' analysis, the pledge functions as a verbal discriminative stimulus, a rule that describes a contingency in which truthful behaviors would have positive consequences and untruthful ones, negative consequences. For public figures, the design provides for reputation as a consequence: volunteer advocates monitor, a committee issues a press release in case of noncompliance, and the figure is listed on the website as being in noncompliance; the article describes the mechanism as a design and reports no actual use or sanctions.

Launched in December 2016, the pledge had more than 7,500 signers by May 2018, more than 1,200 of them public figures. The authors acknowledge that the results of pledges of this kind have been mixed and that false signaling, in which a public figure signs without complying, may prevent the pledge from providing credible information; they hypothesize that the long-term impact will be small or negligible for those who sign without other forms of active en-

gagement. The empirical study assessed past behavior, not current behavior, with 21 participants recruited through Facebook posts and emails to people who had signed and asked for updates. Two coders rated from 1 (the post is misinformation) to 5 (aimed at fighting misinformation and promoting truth) the first ten posts with news content made four weeks after the pledge, compared with the first ten from the same period the previous year. The mean went from 2.49 (SD = 0.60) to 3.65 (SD = 0.41), d = −1.93 (negative because the article subtracts the after score from the before score; the rating rose), with a Krippendorff's alpha of 0.85, an index of agreement between the two coders.

The authors write that the study is evidence that signing the pledge made people behave more honestly on Facebook, and they describe it as preliminary and voluntary, to be replicated with a larger sample and control groups. To this is added what the design imposes: there is no comparison group; the 21 self-selected; the rating mixes the quality of the post with engagement in the comments; and the causal language goes beyond what an uncontrolled pre-post design supports, because those who decide to sign and ask for updates are those who are willing to change. What the study offers that can be used is the description of the arranged contingency, not the effect.

Both studies act on the speaker, after or before the statement, and neither deals with what Chapter 3 identified as the path of false speech in chains: the intraverbal in the form of a tact. Section 3.3 described how the community reinforces the response that has the form of a direct report of the event, even though the speaker emits it on the basis of what he or she heard; Section 3.6, how the source autoclitic is what would indicate that the response is not a tact. Skinner (1957) notes that asking "Did you see it, or did someone tell you?" is asking whether the response was a tact, or else an echoic or intraverbal response to another person's verbal behavior; that in many communities good manners call for opening one's speech with an autoclitic indicating where what is said comes from; and that the community arranges the contingencies for self-descriptive behavior when it has reason to ask "What did you say?," "Did you say that?," and "Why did you say that?" (pp. 314–316).

The hypothesis proposed here leads to an intervention that is not in any study read: teaching and reinforcing, in the community, the discrimination between the tact and the intraverbal in the form of a tact, and the emission of the source autoclitic. It has two arms. The listener's arm consists of asking for the source and not reinforcing, with a like, a repost, or agreement, the statement that arrives without one; the question makes the intraverbal in the form of a tact less reinforced than the statement with its source marked. The speaker's arm consists of marking the source ("I read in...," "I was told that...") and of retracting, which comes close to what Tsipursky et al.'s pledge asks for when it

directs people to verify before sharing and to retract. The point of intervention is the same as in Section 12.3, the moment of reading, with one difference: the stimulus is a consequence arranged by another person, not a printed rule. A design in the format of Hangartner et al. would serve to test it: a neutral profile that, in the face of a claim checked as false, asks where it comes from, versus a control with no reply, measuring the presence of source markers in subsequent posts, deletion of the statement, and reposting, in both political camps.

DATA	Among 1,350 Twitter users, empathy-based counterspeech increased deletion of the tweet by about 8.4 percentage points and reduced, by 0.10 and 0.13 standard deviations, the number of xenophobic tweets and the total, over four weeks, with adjusted P values from 0.053 to 0.112; the share of xenophobic tweets, the preregistered outcome, did not change, and humor and warning had no consistent effect (Hangartner et al., 2021). Among 21 self-selected signers, the rating of Facebook posts went from 2.49 to 3.65, without a control group (Tsipursky et al., 2018).
INTERPRETATION	Both studies measure observed behavior of those who were targeted or who signed, with a four-week follow-up, and neither measures the effect on those who merely read. In the field study, the two components of the share of xenophobic tweets fell by a similar amount, in standard deviations, and the result is compatible with the user tweeting less, not with tweeting less hate. The pledge study is the description of a contingency with an uncontrolled pre-post design.
HYPOTHESIS	A single message from an unknown account may function as mild social punishment, in which case deleting would be escape, and tweeting less, avoidance (reading proposed here, alongside the persuasion the authors propose). Nothing in the data separates the two. As for the source autoclitic, the hypothesis is the section's: the listener who asks for the source alters the consequence of the tact form. There are no data that support it or refute it.

12.5 Reducing prejudice through equivalence and comics

Chapter 4 showed that evaluative relations can be formed and reorganized in the laboratory. The question of this section is a different one: what the sources read say about the use of these procedures to reduce prejudice.

The study by Dixon and Lemke (2007) is the one the literature read cites as a model, and it is worth reading for what it did. Twelve undergraduates (white or African American) at Southern Illinois University participated, for extra credit; the authors operationally define prejudice as differential responding toward members of a specific race. At pretest, each participant rated 30 images (ten of white American men, ten of Middle Eastern men, and ten of everyday objects), from 1 (evil) to 10 (good). The training was matching-to-sample, with differential consequences (a pleasant sound for a correct response, an aversive sound for an incorrect one) and symmetry and equivalence tests in extinction; the trained relation linked the image of a white American man to the word "Evil," that of a Middle Eastern man to the word "Good," and that of an everyday object to the word "Neutral" (p. 10). Four participants rated all the male stimuli "5" at pretest and had their data excluded; the remaining eight rated the Middle Eastern men closer to "Evil" than the American men.

All completed the training, and seven of eight reached 90% or more on each symmetry and equivalence test. The authors write that the Middle Eastern man was now rated closer to "Good" and the American man closer to "Evil," and that the pattern generalized to nine new stimuli. Table 1 of the article allows a count that the authors do not make: in five of the eight participants (1, 3, 5, 8, and 12), the mean for the Middle Eastern man was above that for the American man at posttest, and the mean for the American man fell to 1.00 to 1.50. In the discussion, they say that not all those who reached the equivalence criterion showed significant differences, and they attribute part of the lack of change to implied cultural social contingencies, a greater source of control than the point-based differentiation used in the study; the everyday objects, expected to be rated as neutral, received the highest ratings at pretest and posttest. They caution that participants' responses may not reflect their true relational network, owing to social contingencies that reduce the display of prejudiced behavior, and they note that when equivalence is demonstrated between stimuli with strong pre-existing relations, it may be difficult to alter those relations.

Two points of reading. The first is what the training taught. By the definition in Chapter 4, an attitude in which the evaluative attributes are negative is a prejudice; Dixon and Lemke's procedure related "Evil" to the white American man and "Good" to the Middle Eastern man, so the expected result is not neutralization but a reversal of sign, and the reversal installs in the laboratory a negative evaluation of another group. This is a reading proposed here, and it stands as a caution about design, not as a criticism of the authors, who describe the study as a demonstration. The second is the count: five of eight is a

majority short of unanimity, in a small sample, without a control group and without follow-up. In children, with faces of Black people and a positive symbol, and without a negative class for the other group, Chapter 4 reports that 13 of 13 formed the classes and nine maintained the relations (Mizael et al., 2016a), and that, among 46 children, 33 formed classes and 27 maintained at least one relation after six weeks, with a reduction also in the control group, to a lesser degree (Mizael et al., 2021b). Mizael et al. (2016b) maintain that a training of one or two hours is probably insufficient to modify pre-experimental relations widely reinforced by the verbal community; this is the reviewers' hypothesis.

Comic books come in here because they are the procedure used by the studies of the Instituto Walden4, and for a reason that matters for intervention: the authors of one of the volumes argue that narratives, by treating prejudice subtly, diluted in the plot, minimize the perception of persuasive intent and allow readers to identify with the characters, which traditional methods of attitude change do not achieve in the face of attitudinal resistance (Moreira et al., 2022a). The procedure, the cast, and the scenes are in Figures 4.15 to 4.18 of Chapter 4. The volume edited by Moreira (2025a) brings together single-subject studies, mostly with university students, using matching-to-sample and comics, on sexism, homophobia, ageism, transphobia, and fatphobia, among others, in which attitude is measured by a semantic differential scale; the summary reports that matching-to-sample formed classes and transferred evaluative functions, especially to stimuli that were initially neutral or less emotionally charged, and that there was resistance to changing deeply rooted prejudices or prejudices associated with a strong emotional charge. Experiment 8 of Moreira et al. (2022c) had five adults read a story with three characters, with left-wing, right-wing, and neutral political ideals, and rate the characters at three points in time; the scales changed for all five, for all three characters, in a design in which the subject is his or her own control, with a pretest and a posttest. The criterion applied to the other studies applies to this one: there is no control group, there is no follow-up, the measure is a scale answered right afterward, and the object of the rating is a fictional character, so the change in the ratings cannot, on its own, be distinguished from an effect of repeating the measure (Chapter 4, Section 4.7).

Three reviews situate the whole. Matsuda et al. (2020) searched 14 behavior-analytic journals for keywords related to racism and prejudice in the title (PsycInfo, from 1990), 30 years of literature, and report that the result was zero; broadening the search to conceptual articles and functional-contextual studies, they write that there is a notable deficit in behavior analysis regarding racism and little research to guide the work. They summarize that equival-

ence studies show that bias exists and can be measured, and that established biases are difficult to overcome, although there are promising procedures; they caution that they do not claim that such procedures produce long-term change in prejudice or keep overt racist behavior from occurring outside the laboratory. They cite, as reports that were not read in the original, studies of mindfulness and perspective taking, a pilot ACT-based lesson, and a study with teachers in which, after self-monitoring, performance feedback, and training, children in all categories received equitable rates of praise and reprimands. They conclude that they do not present a definitive solution, because current behavioral research is neither sufficient nor robust, and they recommend extending and replicating studies, studying longitudinal outcomes, and developing community-wide interventions.

Shea et al. (2023) review, in a narrative article with no search method described, the common strategies against prejudice (education, diversity training, and shaming) through Relational Frame Theory, because their results are limited. They report, from works that were not read in the original, small, self-reported changes in attitudes, which decayed over time, in a meta-analysis of 230 samples over 40 years; a harassment prevention training that slightly worsened the attitudes of 90 male undergraduates; a diversity training literature that is fragmented and limited by poor measures and a lack of maintenance and behavioral data; and a shaming procedure with no effect on attributions related to disabilities. The authors' hypothesis is that the method of delivering education may strengthen prejudice by re-presenting the prejudiced views, and that shaming may function as a reflexive conditioned motivating operation that leads people to hide prejudice. They conclude that research in Relational Frame Theory, ACT, and IRAP has shown promising outcomes, but that longitudinal research is essential and that treatments must not perpetuate prejudiced relational networks; they recommend, as a hypothesis, multiple exemplar training with marginalized groups before children are exposed to networks that promote prejudice.

Levy et al. (2022) add a methodological constraint. They say that the behavior-analytic literature often fails to report sufficient demographic variables, including variables related to race and to culturally diverse populations and settings, which may lead to a poor understanding of the efficacy of measures and assessments. The studies in this section generally report little about who the participants are and for whom the result holds.

DATA	In the laboratory, equivalence training changed ratings in the trained direction in five of eight adults (our own count from Dixon and Lemke, 2007), in 13 of 13 children with maintenance in nine (Mizael et al., 2016a), and in 33 of 46 with partial maintenance at six weeks (Mizael et al., 2021b). A search of 14 journals found no studies with the terms racism or prejudice in the title (Matsuda et al., 2020), and the reviews say there are no data on long-term effects or on overt behavior outside the laboratory (Matsuda et al., 2020; Shea et al., 2023).
INTERPRETATION	Evaluative relations are trainable and, in part, are maintained for weeks, in small samples and with scale measures. Nothing that was read shows a change in what people say or do in other contexts, or in what the community reinforces.
HYPOTHESIS	The old relations continue to be reinforced by the verbal community outside the procedure, and that is why one-off training tends to be insufficient, as Mizael et al. (2016b) and Shea et al. (2023) maintain (the authors' hypotheses). Two readings follow from this: that the intervention that changes only the relation in the individual, without touching what the community reinforces, is the one the sources least support; and that the form of delivery (correcting, instructing, shaming) may re-present the relation one wants to undo, as Section 12.1 raised for correction.

12.6 Changing platform contingencies

This section deals with the platform level. Lazer et al. (2018) discuss two broad mitigation strategies: helping individuals become better at identifying fake news, and intervening within platforms to reduce the attention it receives; and they point out that platform cooperation with researchers is critical to sizing the problem and the effectiveness of interventions. The previous sections dealt with the first. This one deals with the second, which acts on what Chapter 8 described as the delivery rule and Chapter 11 called reach. It is also the one most dependent on access: the studies read are few, and most were conducted with the cooperation of the platform itself.

What the delivery rule does was measured by Huszár et al. (2022), on Twitter. With the personalization of the timeline, introduced in 2016, all parties had positive amplification (the ratio between reach in the group with the personalized timeline and in the control group, which kept the chronological one; 0% is equal reach), in some cases above 200%, that is, an audience three times

that of the chronological timeline; with the exception of Germany, the largest right-wing or center-right party had greater amplification than the left-wing or center-left one, and, at the individual level, a permutation test did not detect a statistically significant association between a politician's party and his or her amplification. This is not an evaluated intervention but the measurement of a rule that was already in operation. The control group, of 1% of users, was not created for research, and the authors themselves say that it is not isolated from indirect effects, so the statistics probably underestimate the effects; the work went through review by Twitter's legal and privacy teams, not an ethics committee.

Nyhan et al. (2023) did what Huszár et al. could not: a randomized intervention in the feed. The 23,377 participants, U.S. adults, were recruited through invitations at the top of their Facebook feeds, gave consent, and answered surveys in waves before, during, and after the election. For those assigned to treatment, all content from friends, groups, and pages predicted to share the participant's political leaning was demoted in the feed, from September 24 to December 23, 2020, with the greatest demotion strength that a pretest indicated would not leave the feed nearly empty; the control group's feed was not systematically altered. Exposure to like-minded sources fell from 53.7% to 36.2%; exposure to cross-cutting sources rose from 20.7% to 27.9%, and to sources that were neither like-minded nor cross-cutting, from 25.6% to 35.9%. There was no measurable effect on eight preregistered attitude measures, such as affective polarization, ideological extremity, candidate evaluations, and belief in false claims, and the authors rule out effects of ±0.12 standard deviations or more, that is, of just over a tenth of the typical variation of each measure. Those whose exposure was reduced engaged more, per exposure, with the like-minded content they still saw. Among the possible explanations of the null result, the authors cite that congenial political content is a fraction of what people see, that Facebook is a fraction of the information diet, that persuasion is difficult, and that the design does not capture effects of prior use or effects accumulated over years; they also note that reducing exposure could lead people to seek that information elsewhere on the internet.

The third study deals with moderation. Álvarez-Benjumea and Winter (2018) asked 180 participants, recruited on a crowdsourcing platform, to comment, in a simulated forum, on nine photos on controversial topics (such as feminism, LGBT rights, and refugees), in four conditions: baseline (two friendly comments, two neutral ones, and two hostile ones), censored (hateful content deleted), extremely censored (only friendly comments), and counterspeech (hostile comments with replies highlighting their unacceptability). Hostile comments (rating of 7 or more) numbered 24 (5.88%) in the baseline, 4 (0.99%)

in the censored condition, 18 (4.44%) in counterspeech, and 15 (4.10%) in the extremely censored condition; only the reduction in the censored condition was significant. The authors suggest that normative online behavior is driven by descriptive norms, what others do, more than by injunctive norms, and that the replies, being themselves hostile, may add ambiguity about the norm. They caution that they have no data on the trade-off between censorship and freedom of expression and that the anonymity of the forum precludes group identification processes.

Ceylan et al. (2023), whom Chapter 9 read, come in here for two reasons. The first is study 4, which works as an experimental model of a delivery rule: in 80 trials, participants earned "+5 points" when their choice followed the training criterion (in accuracy training, sharing a true headline or not sharing a false one; in misinformation training, the reverse) or received no feedback (control); in the test, without feedback, the pattern persisted. After accuracy training, 66% of the true headlines and 24% of the false ones were shared; in the control condition, 40% and 24%; after misinformation training, 54% and 33%. What the feedback did was, in the first case, increase the sharing of the true without reducing that of the false, and, in the second, increase sharing in general. The second reason is the authors' conclusion: habitual users are an integral part of social media's advertising-based profit models and, for that reason, platforms are unlikely to create reward structures that prevent habits.

DATA	Timeline personalization amplified all parties, with more amplification on the right than on the left among the largest parties in six of the seven countries and with no detectable association at the level of the individual politician (Huszár et al., 2022). Demoting like-minded sources on Facebook for three months reduced that exposure from 53.7% to 36.2% and had no measurable effect on eight attitude measures (Nyhan et al., 2023). Deleting hostile comments in a simulated forum reduced hostile comments from 5.88% to 0.99% (Álvarez-Benjumea & Winter, 2018).
INTERPRETATION	The delivery rule changes what people see, and moderation changes what is posted in a simulated forum; but, in the only randomized feed experiment read, a reduction of 17.5 percentage points in exposure to like-minded sources did not change the attitudes measured. Exposure and attitude are different variables (Chapter 11).

HYPOTHESIS What the delivery rule alters is the occasion (reach) and, through the audience's reaction, the consequence for the producer (Chapter 9). If so, the intervention that reaches what maintains production is the one that changes the feedback the producer receives, and no study read has tested it: all measured what the audience sees or does, and none what those who post went on to post. Testing it would require varying the feedback (reach or visible reaction) and measuring the content posted afterward, with the same criterion for producers from any political camp.

Table 12.1 brings together the eleven interventions from sections 12.1 to 12.6, with the design, the measure, the result, and the limitation that the study itself states.

TABLE 12.1

The interventions read: design, measure, result, and limitation

Study	Intervention and design	Measure	Result	Limitation
Nyhan and Reifler (2010)	Correction embedded in a mock news article; four between-subjects experiments, students	Agreement with the claim, right after reading	Backfire in two of four, among conservatives; the correction worked in the other two	Backfire only among conservatives; no follow-up; student sample
Freeze et al. (2021)	Misinformation warning, invalid or valid; online experiment, 434 participants	Credibility and memory	Warning reduced credibility by 0.59 points (out of 5), even for an accurate article; valid warning, 46% to 52% memory ($p = 0.06$)	General warning, from the researcher; session only
Roozenbeek and van der Linden (2019)	The game Bad News; pre-post without control, 14,266 pairs	Reliability from 1 to 7	d from 0.30 to 0.36 for three techniques; real items unchanged	No control; self-selected, 75% men; no follow-up
Roozenbeek et al. (2022)	Inoculation videos; six randomized experiments and one field study (22,632)	Technique recognition, trustworthiness, willingness to share	Hypotheses supported for 16 of 20 measures; field, $h = 0.09$	Within 24 hours; United States; susceptibility, not behavior
Guess et al. (2020)	Facebook's list of tips; two-wave panel (U.S. and India) and face-to-face sample	Perceived accuracy, intention to share, observed consumption	U.S.: discrimination +26.5%; effect falls by more than half in 20 days; India face-to-face, no effect	No evidence for intention or consumption

Study	Intervention and design	Measure	Result	Limitation
Dixon and Lemke (2007)	Equivalence training with reversed valence; 12 students, 8 analyzed	Ratings of images from 1 to 10	Five of eight reversed the order (our own count)	No control or follow-up; objects received the highest ratings
Mizael et al. (2021b)	Training with different parameters; 46 children; six-week follow-up	Class formation and maintenance; bias	33 of 46 formed classes; 27 maintained at least one relation	Association, not cause; control group also reduced bias
Hangartner et al. (2021)	One-shot counterspeech by a neutral profile; 1,350 Twitter users; four weeks	Xenophobic tweets created and deleted	Empathy: deletion +0.21 SD (adjusted, 0.053); xenophobic tweets −0.10 SD and total −0.13 SD (adjusted, 0.112); preregistered outcome with no effect	Does not measure those who merely read; humor and warning with no effect
Tsipursky et al. (2018)	Public commitment (Pro-Truth Pledge); pre-post, 21 signers	Rating by two coders, from 1 to 5, of ten posts	2.49 to 3.65 (d = −1.93)	No control; self-selection; past behavior
Álvarez-Benjumea and Winter (2018)	Removal of hostile comments; simulated forum; 180 participants	Hostile comments (rating of 7 or more)	5.88% in the baseline and 0.99% in the censored condition; counterspeech with no effect	No data on the trade-off with freedom of expression
Nyhan et al. (2023)	Demoting like-minded sources in the Facebook feed; three months; n = 23,377	Exposure and eight attitude measures	Exposure from 53.7% to 36.2%; no effect on attitudes (±0.12 SD ruled out)	Accumulated effects not captured; United States, 2020

Note. Our own elaboration, based on the articles; the figures and pages are in sections 12.1 to 12.6. The "Limitation" column brings together the authors' caveats and, in some cases, what the design imposes and the text discusses; our own counts are flagged.

12.7 Cultural design and the ethical limits of those who design

Of the chapter's four levels of intervention, culture is the only one at which no study read evaluated an intervention, the empty row of Figure 12.1. This is not for lack of a concept. Glenn (2004) addresses, in the final pages of her article, what can be designed when the object is a practice and not a person, and the concept the analysis needs is hers.

The reading allows us to reorder what the chapter has seen. The interventions in sections 12.1 to 12.3 and 12.5 act on each individual's occasion, by showing, teaching, or warning, and they assume, in Glenn's vocabulary, that changing the behavior of many people individually would change the cumulative effect;

those in Section 12.4 act on the consequence that one person arranges for another; those in 12.6 act on the environment that selects what circulates. The hypothesis of Chapters 8 and 9 is that what maintains speech lies in the consequence, not in the occasion. If so, the evidence is where measurement is cheap, and the hypotheses are where what maintains speech is. This is the chapter's central observation, and it is an interpretation: nothing in the data shows that intervention on the consequence works better, because no study read has tested it.

Levy et al. (2022) give design a condition that the other sources do not. They describe cultural humility as a value that guides interaction with people from other cultural backgrounds, a continuing pattern of behavior, not a terminal outcome that can be reached, and they criticize "cultural competence" for implying that learning has an endpoint. They write that the behaviors that constitute racism are taught, imitated, and maintained, incidentally or through deliberately arranged contingencies, and that racism exists at the individual, systemic, and cultural levels. They mention the metacontingencies selected to maintain racist systems and say that examining them is outside the scope of the article. And they say that behavior analysis, given its Eurocentric roots and the lack of diversity, equity, and inclusion observed in it, is not exempt from evaluation, and that it needs to be adapted to the cultural variability of its members and of the people it serves. Reading proposed here: those who design are part of the practice they design, and so is the science that offers the tools.

CONCEPT IN USE

Cultural design (Glenn, 2004)

Glenn writes that cultural engineering still lacks a secure scientific foundation, that it must begin by understanding the phenomena it intends to engineer, and that it can target two kinds of phenomena: macrobehavior and metacontingencies. The only way to act on the cumulative effects of macrobehavior is to change the behavior of the largest possible number of individual participants; when that number is large, changing a small percentage of them can make a considerable difference. To design interventions, one must identify which operant contingencies explain the behavior of the individuals taking part in the practice, and this implies changing the operant contingencies that maintain the behavior of many people. Recurring interlocking behavioral contingencies can be changed in two ways: by altering the external selecting environment and waiting for variations to arise, or by altering their components so that they fit the current selecting environment better. For Glenn, without an understanding of how cultures emerge and change, choosing wisely what can or should be changed will be hard. Reading proposed here: cultural design is the set of interventions that change the operant contingencies of many participants or the environment that selects practices, not the information each person receives.

Hence the ethical limits, which the chapter gathers into five, all drawn from the sources and from the criterion adopted here. The first is who decides what is false or what is classified as hate speech. Chapter 2 showed that the label of false depends on who checks, with what evidence, and at what moment, and Chapter 5 that "hate speech" enters this book only as an external label; designing the reduction of something defined by a third party makes the third party part of the design. The second is freedom of expression. Zielinski (2021) warns that corrective actions need human rights safeguards so as not to lead to censorship and the restriction of freedom of expression, and that technology alone does not solve the problem; Hangartner et al. (2021) start from the premise that moderation may suppress valuable speech and disperse hate, and Álvarez-Benjumea and Winter (2018) say they have no data on the trade-off between censorship and freedom of expression.

The third limit is that of the effect that was not sought. Hangartner et al.'s empathy message also reduced the total number of tweets, and the authors call for attention to chilling effects beyond hate speech; Freeze et al.'s general warning reduced the credibility of accurate news; Dixon and Lemke's equivalence training reversed the valence; Nyhan and Reifler's correction raised, among conservatives, agreement with the claim in two of the four experiments. Every intervention must also be read for what it installs alongside what it removes. The fourth is that of consent and control by third parties.

Huszár et al.'s work was reviewed by the company's legal and privacy teams in the normal course of its business, rather than by an ethics committee, and a data protection impact assessment found no need for further notice and consent mechanisms. Hangartner et al.'s was approved by an ethics committee and preregistered, and the intervention came from profiles created for the study, with a neutral and apolitical history; the text read makes no mention of user consent. No decision is made here between the two approaches; it is recorded that, in both, the decision about notice and consent fell to those who conducted or reviewed the study.

The fifth is that of the symmetric criterion. What Chapter 10 established for the Overton window applies here: the criterion is the same for what widens and for what narrows, for the political camp of whoever emits, for the language, and for the platform, and the caveat that each study measures what it measures is the same. An intervention whose effects are measured only for the content of one camp does not meet the criterion, whatever the camp.

For those who wish to use the chapter in reading an intervention, a news story about one, or a platform report, the questions the sections suggest are the following.

1. At what level does it act (individual, interaction, platform, or culture), and on which term of the contingency (occasion, response, or consequence)?
2. What was measured: judgment, stated intention, or observed behavior? With what follow-up, and in whom (those who read, those who speak, or those who post)?
3. What was the comparison, and was assignment random? What do the authors say the design does not allow them to claim?
4. Does the result hold for content from another political camp, in another language, and on another platform? Was the criterion the same for all?
5. What does the intervention install or teach beyond the target: a signal without a target, a reversed valence, a rule that competes with the audience's reaction?
6. Who decides what counts as false or as hate speech, and what safeguard is there against the restriction of expression?

Summary

The chapter organized interventions into four levels and read eleven studies for what each one measured. Eight measured judgment or task performance, five measured observed behavior, four, behavior with a follow-up of one week

or more, and none evaluated an intervention at the cultural level (Figure 12.1). Correction after reading has an uncertain effect and depends on the topic; Nyhan and Reifler's backfire came from misperceptions held by conservatives, on the topics chosen, and the general warning reduced the credibility of accurate news (12.1).

Inoculation, read as discrimination training with exemplars of the technique, improves, shortly afterward, the rating of new items that use it, with a small effect in the field and no measure of behavior (12.2). The tips for spotting false news changed judgment for a short time, where they were understood, and did not change, in the data, consumption, nor, with sufficient evidence, the stated intention to share (12.3). In the face of the speaker, empathy-based counterspeech increased deletion and reduced activity, with no effect on the preregistered outcome, and the public pledge has an uncontrolled pre-post design (12.4); the hypothesis proposed here, that of teaching and reinforcing the source autoclitic, has not been tested.

Equivalence training changes ratings in the laboratory, in part and for weeks, and no source read shows a change in what is said or done outside it; the reviews call for follow-up and community intervention, and one of them cautions that re-presenting the relation may strengthen it (12.5). The delivery rule changes what is seen, and demoting like-minded sources did not change the attitudes measured; moderation reduced hostile comments in a simulated forum, and no study measured what those who post went on to post (12.6). For Glenn, acting on cumulative effects requires changing the operant contingencies of many participants or the environment that selects practices; the ethical limits are who defines what is false, freedom of expression, side effects, consent, and the symmetry of the criterion (12.7).

What the chapter leaves is an asymmetry. The evidence is where measurement is cheap, in a person's judgment in front of a screen, and the hypotheses are where, according to the analysis in Chapters 8 and 9, what maintains speech lies: in the consequence that the audience and the platform arrange for those who produce it. The question that remains for researchers is that of a design that measures the behavior of those who post, in the face of a change in feedback, with follow-up and with the same criterion for all camps.

References

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Calleja, N., AbdAllah, A., Abad, N., Ahmed, N., Albarracin, D., Altieri, E., Anoko, J. N., Arcos, R., Azlan, A. A., Bayer, J., Bechmann, A., Bezbaruah, S., Briand, S. C., Brooks, I., Bucci, L. M., Burzo, S., Czerniak, C., De Domenico, M., Dunn, A. G., . . . Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference. *JMIR Infodemiology*, *1*(1), Article e30979. https://doi.org/10.2196/30979

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Dixon, M. R., & Lemke, M. (2007). Reducing prejudice towards Middle Eastern persons as terrorists. *European Journal of Behavior Analysis*, *8*(1), 5–12. https://doi.org/10.1080/15021149.2007.11434269

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues*, *12*(2), 129–147. https://doi.org/10.5210/bsi.v12i2.40

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Glenn, S. S. (2004). Individual behavior, culture, and social change. *The Behavior Analyst*, *27*(2), 133–151. https://doi.org/10.1007/bf03393175

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), Article e2025334119. https://doi.org/10.1073/pnas.2025334119

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Levy, S., Siebold, A., Vaidya, J., Truchon, M.-M., Dettmering, J., & Mittelman, C. (2022). A look in the mirror: How the field of behavior analysis can become anti-racist. *Behavior Analysis in Practice*, *15*(4), 1112–1125. https://doi.org/10.1007/s40617-021-00630-3

Matsuda, K., Garcia, Y., Catagnus, R., & Brandt, J. A. (2020). Can behavior analysis help us understand and reduce racism? A review of the current literature. *Behavior Analysis in Practice*, *13*(2), 336–347. https://doi.org/10.1007/s40617-020-00411-4

Mizael, T. M., de Almeida, J. H., Roche, B., & de Rose, J. C. (2021b). Effectiveness of different training and testing parameters on the formation and maintenance of equivalence classes: Investigating prejudiced racial attitudes. *The Psychological Record*, *71*(2), 265–277. https://doi.org/10.1007/s40732-020-00435-w

Mizael, T. M., de Almeida, J. H., Silveira, C. C., & de Rose, J. C. (2016a). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record*, *66*(3), 451–462. https://doi.org/10.1007/s40732-016-0185-0

Mizael, T. M., Dos Santos, S. L., & De Rose, J. C. C. (2016b). Contribuições do paradigma de equivalência de estímulos para o estudo das atitudes [Contributions of the stimulus equivalence paradigm to the study of attitudes]. *Interação em Psicologia*, *20*(2). https://doi.org/10.5380/psi.v20i2.46278

Moreira, M. B. (Ed.). (2025a). *Formação e mudança de atitudes: Pareamento ao modelo e história em quadrinhos* [Attitude formation and change: Matching-to-sample and comics] (1st ed.). Instituto Walden4.

Moreira, M. B., Premuli, I. T., Rabelo, E. de P., Pessoa, M. B. F., & Medeiros, M. R. G. C. (2022a). *Atitudes: Histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos* [Attitudes: Comics, semantic analyses on social media, and the stimulus equivalence paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_q0.pdf

Moreira, M. B., Sousa, C. P., Tomasini, C., Craveiro, H. do A., Rodrigues, I. L., Dutra, I., Barros, J. F., Freire, L. R. V., D'arcanchy e Vilela, M. E., Cautela, M. E. T. C., Oliveira, P. M. D. N., Soares, S. K. S., & Andrade, V. M. (2022c). *Estudos sobre histórias em quadrinhos e atitudes - Parte I* [Studies on comics and attitudes, Part I] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_hq1.pdf

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., . . . Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137–144. https://doi.org/10.1038/s41586-023-06297-w

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Shea, P., Johnson, P., & Togade, D. (2023). Using relational frame theory to examine racial prejudice: A tool for educators and an appeal for future research. *Behavior Analysis in Practice*, *16*(1), 102–116. https://doi.org/10.1007/s40617-022-00767-9

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Tsipursky, G., Votta, F., & Roose, K. M. (2018). Fighting fake news and post-truth politics with behavioral science: The Pro-Truth Pledge. *Behavior and Social Issues*, *27*(1), 47–70. https://doi.org/10.5210/bsi.v27i0.9127

Zielinski, C. (2021). Infodemics and infodemiology: A short history, a long future. *Revista Panamericana de Salud Pública*, *45*, 1. https://doi.org/10.26633/rpsp.2021.40

APPENDIX

A

Discourse Analysis Form

This is the form from Chapter 3 (Section 3.2), ready for use. Its purpose is to force a separation between what is known and what is inferred: the first five fields are data, the next three are interpretation, and the last two are hypothesis. Where the source does not document something, write "not documented"; a field is never filled in by inference. The same form, with the same criterion, applies to statements from any political camp.

#	Field	What to record	Nature
1	Statement	The utterance, transcribed as the source gives it, in quotation marks, with the source reference	Data
2	Date	Date of the statement	Data
3	Context and outlet	Where and in what outlet it was said; the question, argument, or event immediately preceding it, if the source reports them	Data
4	Audience present	Who was present and how they reacted, when the source documents it	Data
5	Fact-check	Agency, label, date, and rationale; the falsity always belongs to the agency that classified it (Chapter 2)	Data
6	Probable operant	Mand, tact, intraverbal, distorted tact, disguised mand or impure tact, ritual or identity function; and why (Tables 3.1 and 3.4)	Interpretation
7	Candidate controlling variables	Occasion, audience, probable consequence, reinforcement history, what the listener would do next	Interpretation
8	Formal markers	Source and certainty autoclitics, quantifiers, repetition (verbatim or with variation)	Interpretation

#	Field	What to record	Nature
9	What would change the statement	The observation that would decide among the readings: the occasion, the audience, the comparison across audiences or before and after the correction	Hypothesis
10	Rival interpretations	The six in Table 3.4, with what would distinguish each one: weak control by the event, intraverbal acquired from others, distorted tact, disguised mand or impure tact, ritual or identity function, effect of selection by the agency	Hypothesis

The form calls for four kinds of observation that records made by third parties usually lack: the occasion (what the statement responds to); the audience (who was present and how they reacted); the comparison of the same speaker and the same event before different audiences, or before and after the fact-check; and what happens after the correction (whether the form persists, varies, or disappears). Without them, the analysis remains interpretation.

Blank form

Field	Record
1. Statement	
2. Date	
3. Context and outlet	
4. Audience present	
5. Fact-check (agency, label, date)	
6. Probable operant	
7. Candidate controlling variables	
8. Formal markers	
9. What would change the statement and how one would know	
10. Rival interpretations and what distinguishes them	
Not documented	

APPENDIX

B

Concepts in Use

Each concept is presented once, in a "Concept in use" box (or "Proposed definition," when the term is proposed here), in the first chapter whose analysis depends on it; later chapters return to it and refer back to the original box. The table follows the order of the book.

Concept	Chapter (section)	Short definition	Main source
Verbal behavior, audience, and self-report	1 (1.5)	Behavior reinforced through the mediation of other people; the audience is the discriminative stimulus in whose presence it is characteristically reinforced. Self-report (scale, question, interview) is verbal behavior, under the control of who asks and how they ask	Skinner (1957); Guerin (1994)
Belief and attitude as verbal behavior	2 (2.8)	Attitudes and beliefs are behavior shaped by verbal communities, not reports of an internal state; a belief can be contested with evidence, an attitude not in the same way	Guerin (1994)
Autoclitic	2 (2.8); 3 (3.6)	A qualification added to a statement that modifies its effect on the listener	Guerin (1994); Skinner (1957)
Three-term contingency, function × topography, and the verbal episode	3 (3.1)	A verbal operant is defined by what controls the response, not by its form; the unit is occasion, response, and consequence; the episode includes speaker and listener	Skinner (1957)
Generalized social reinforcement	3 (3.3); revisited in 8 (8.1) and 9 (9.1)	An event that usually precedes many different reinforcers and reinforces behavior under various conditions of deprivation; approval is the example. For Guerin, attitudes are statements maintained by it	Skinner (1957); Guerin (1994)

Concept	Chapter (section)	Short definition	Main source
Distorted tact and verbal subcommunity	3 (3.4); revisited in 6 (6.1)	A tact whose control by the event has been displaced by special consequences from the listener; a subcommunity smaller than the broader one can control what is reported	Skinner (1957); Guerin (1992)
Source and certainty autoclitics	3 (3.6)	Autoclitics that indicate the source or the strength of the response; quantifiers (“all,” “none”) modify the listener’s reaction to the whole sentence	Skinner (1957)
Rule and rule-governed behavior	3 (3.7)	A verbal discriminative stimulus that describes a contingency; instructed performance is less sensitive to changes in consequences	Baum (1995); Cerutti (1989); Hayes et al. (1986)
Attitude	4 (4.1)	The verbal response of attributing to an object a value ranging from positive to negative, maintained by the consequences the verbal community arranges; the object need not have been contacted	Guerin (1994); Cameoka and Moreira (2021)
Conditional discrimination and matching-to-sample	4 (4.2)	Given a sample and comparisons, the choice is consequated according to the sample: if A1, then B1	Sidman and Tailby (1982)
Equivalence class	4 (4.2)	A set of stimuli whose conditional relations show reflexivity, symmetry, and transitivity	Sidman and Tailby (1982)
Transfer (or transformation) of function	4 (4.3)	A function acquired by one stimulus through direct contact appears in the other members of the class without additional training; that it “transferred” is an interpretation	Dougher et al. (1994); Dymond and Rehfeldt (2000)
Derived relations of opposition and comparison	4 (4.6)	Patterns of relational responding under contextual control: in opposition, A opposite to Y and to Z implies Y same as Z; in comparison, the relation defines how the function is transformed	Dixon et al. (2003); Weinstein et al. (2008)
Interlocking behavioral contingencies	5 (5.1); revisited in 9 (9.4)	The behavior of two or more people is an environmental event for that of the others; the outcome is that of the interrelated behavior	Glenn (2004)
Disguised mand	5 (5.2)	A statement with the form of a tact or an opinion whose function, on the available evidence, is that of a mand; its application to political speech is a hypothesis proposed here	Skinner (1957); Guerin (1994)

Concept	Chapter (section)	Short definition	Main source
Aversive control	5 (5.3)	Control of behavior by aversive stimulation (escape, avoidance, punishment); it can be verbal and mild	Skinner (1957); Fazzano and Gallo (2015)
Countercontrol	6 (6.4)	Episodes of socially mediated aversive controlling conditions and of escape or avoidance responses that do not reinforce the controlling behavior	Delprato (2002)
Avoidance of countercontrol	6 (6.4)	A term proposed here, as a hypothesis: a repertoire of the speaker or of the community maintained because it reduces the probability or the effectiveness of countercontrol	Proposed here, based on Delprato (2002)
Say–do correspondence	7 (7.4)	The relation between what a person says they did or will do and what they do; these are separate classes, and correspondence is an effect of contingency	Risley and Hart (1968); Lloyd (1994)
Attention as a generalized conditioned reinforcer	8 (8.1)	The listener's attention is a common consequence of generalized mands and reinforces verbal behavior regardless of form	Skinner (1957)
Intermittent reinforcement	8 (8.4)	Only some of the responses or occasions are followed by the reinforcer; in verbal behavior, intermittency is the rule	Skinner (1957)
Schedules of reinforcement	8 (8.4)	The reinforcer depends on the time since the previous one (interval) or on the number of responses (ratio), fixed or variable	Ferster and Skinner (1957)
Metacontingency and macrocontingency	9 (9.7); macrocontingency revisited in 11 (11.4)	Metacontingency: a cultural practice and the aggregate outcome of the current variations, which selects interlocking contingencies. Macrocontingency: the relation between the practice and the aggregate sum of the consequences, which does not control the individual	Glenn (1988, 2004); Houmanfar and Rodrigues (2006)
Negative audience and nonpunishing audience	10 (10.1)	Negative audience: in its presence verbal behavior is punished; nonpunishing: it merely does not reinforce. Two audiences together create a dilemma	Skinner (1957)
An audience's window	10 (10.1)	Proposed definition: for a given speaker and topic, the statements the speaker emits before that audience without being punished by it	Proposed here
Reach, exposure, and reaction	11 (11.2)	Proposed definition: reach is who receives or can see the emission; exposure, what they actually saw; reaction, what they do afterward	Proposed here

Concept	Chapter (section)	Short definition	Main source
Discrimination training	12 (12.2)	The prior stimulus becomes the occasion for the response through operant discrimination; with multiple exemplars, the response comes under the control of the common property and generalizes	Skinner (1957); Shea et al. (2023)
Cultural design	12 (12.7)	Interventions that change the operant contingencies of many participants or the environment that selects practices (macrobehavior and metacontingencies)	Glenn (2004)

APPENDIX

C

References

The works cited in the chapters, in alphabetical order (APA 7). When authors and year coincide, the letter following the year is the one used in the in-text citations. All of them were read in full text, in the version indicated in Appendix D (Section D.3).

Agência Lupa. (2026, July 27). *Anatomia da desinformação eleitoral* [Anatomy of electoral disinformation] [Report]. https://uploads.agencialupa.org/2026/07/RELATORIO-ANATOMIA-DA-DESINFORMACAO-ELEITORAL_LUPA-FINAL.pdf

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52*, 27–58. https://doi.org/10.1146/annurev.psych.52.1.27

Ajzenman, N., Cavalcanti, T., & Da Mata, D. (2023). More than words: Leaders' speech and risky behavior during a pandemic. *American Economic Journal: Economic Policy*, *15*(3), 351–371. https://doi.org/10.1257/pol.20210284

Albarracin, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, *69*, 299–327. https://doi.org/10.1146/annurev-psych-122216-011911

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Barnes, D., Lawlor, H., Smeets, P. M., & Roche, B. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record*, *46*(1), 87–107. https://doi.org/10.1007/bf03395165

Baum, W. M. (1995). Rules, culture, and fitness. *The Behavior Analyst*, *18*(1), 1–21. https://doi.org/10.1007/bf03392688

Bohner, G., & Dickel, N. (2011). Attitudes and attitude change. *Annual Review of Psychology*, *62*, 391–417. https://doi.org/10.1146/annurev.psych.121208.131609

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Bonow Soares, F., Recuero, R., Volcan, T., Fagundes, G., & Sodré, G. (2021). Research note: Bolsonaro's firehose: How Covid-19 disinformation on WhatsApp was used to fight a government political crisis in Brazil. *Harvard Kennedy School Misinformation Review*, *2*(1). https://doi.org/10.37016/mr-2020-54

Bortoloti, R., & de Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record*, *59*(4), 563–590. https://doi.org/10.1007/bf03395682

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), Article eabe5641. https://doi.org/10.1126/sciadv.abe5641

Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, *110*(11), 3522–3548. https://doi.org/10.1257/aer.20171175

Calleja, N., AbdAllah, A., Abad, N., Ahmed, N., Albarracin, D., Altieri, E., Anoko, J. N., Arcos, R., Azlan, A. A., Bayer, J., Bechmann, A., Bezbaruah, S., Briand, S. C., Brooks, I., Bucci, L. M., Burzo, S., Czerniak, C., De Domenico, M., Dunn, A. G., . . . Purnat, T. D. (2021). A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference. *JMIR Infodemiology*, *1*(1), Article e30979. https://doi.org/10.2196/30979

Cameoka, M. C., & Moreira, M. B. (2021). Preconceito racial: viés na mensuração de atitudes produzido por controle de estímulos [Racial prejudice: Bias in attitude measurement produced by stimulus control]. *Acta Comportamentalia*, *29*(1), 93–112. https://doi.org/10.32870/ac.v29i1.78782

Carvalho, P. R., Sousa, P. C. C., & Schneider, M. A. F. (2021). Desinformação na pandemia de Covid-19: similitudes informacionais entre Trump e Bolsonaro [Disinformation in the Covid-19 pandemic: Informational similarities between Trump and Bolsonaro]. *Em Questão*, *27*(3), 15–41. https://doi.org/10.19132/1808-5245273.15-41

Catania, A. C., Matthews, B. A., & Shimoff, E. (1982). Instructed versus shaped human verbal behavior: Interactions with nonverbal responding. *Journal of the Experimental Analysis of Behavior*, *38*(3), 233–248. https://doi.org/10.1901/jeab.1982.38-233

Cerutti, D. T. (1989). Discrimination theory of rule-governed behavior. *Journal of the Experimental Analysis of Behavior*, *51*(2), 259–276. https://doi.org/10.1901/jeab.1989.51-259

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), Article 16598. https://doi.org/10.1038/s41598-020-73510-5

de Carvalho, M. P., & de Rose, J. C. (2014). Understanding racial attitudes through the stimulus equivalence paradigm. *The Psychological Record*, *64*(3), 527–536. https://doi.org/10.1007/s40732-014-0049-4

de Rose, J. C., McIlvane, W. J., Dube, W. V., Galpin, V. C., & Stoddard, L. T. (1988). Emergent simple discrimination established by indirect relation to differential consequences. *Journal of the Experimental Analysis of Behavior*, *50*(1), 1–20. https://doi.org/10.1901/jeab.1988.50-1

DeFleur, M. L., & Westie, F. R. (1958). Verbal attitudes and overt acts: An experiment on the salience of attitudes. *American Sociological Review*, *23*(6), 667–673. https://doi.org/10.2307/2089055

Delprato, D. J. (2002). Countercontrol in behavior analysis. *The Behavior Analyst*, *25*(2), 191–200. https://doi.org/10.1007/bf03392057

Dixon, M. R., & Lemke, M. (2007). Reducing prejudice towards Middle Eastern persons as terrorists. *European Journal of Behavior Analysis*, *8*(1), 5–12. https://doi.org/10.1080/15021149.2007.11434269

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues*, *12*(2), 129–147. https://doi.org/10.5210/bsi.v12i2.40

Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior*, *62*(3), 331–351. https://doi.org/10.1901/jeab.1994.62-331

Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst*, *23*(2), 239–254. https://doi.org/10.1007/bf03392013

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Evangelista, R., & Bruno, F. (2019). WhatsApp and political instability in Brazil: Targeted messages and political radicalisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1434

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, *11*(1), Article e11. https://doi.org/10.2196/jmir.1157

Eysenbach, G. (2020). How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research*, *22*(6), Article e21820. https://doi.org/10.2196/21820

Fazzano, L. H., & Gallo, A. E. (2015). Uma análise da homofobia sob a perspectiva da análise do comportamento [An analysis of homophobia from the perspective of behavior analysis]. *Temas em Psicologia*, *23*(3), 535–545. https://doi.org/10.9788/tp2015.3-02

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. Appleton-Century-Crofts. https://doi.org/10.1037/10627-000

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, *14*(2), 259–269. https://doi.org/10.1177/19485506221083811

Galizio, M. (1979). Contingency-shaped and rule-governed behavior: Instructional control of human loss avoidance. *Journal of the Experimental Analysis of Behavior*, *31*(1), 53–70. https://doi.org/10.1901/jeab.1979.31-53

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), Article eabf1234. https://doi.org/10.1126/sciadv.abf1234

Glenn, S. S. (1988). Contingencies and metacontingencies: Toward a synthesis of behavior analysis and cultural materialism. *The Behavior Analyst*, *11*(2), 161–179. https://doi.org/10.1007/bf03392470

Glenn, S. S. (2004). Individual behavior, culture, and social change. *The Behavior Analyst*, *27*(2), 133–151. https://doi.org/10.1007/bf03393175

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158

Guerin, B. (1992). Behavior analysis and the social construction of knowledge. *American Psychologist*, *47*(11), 1423–1432. https://doi.org/10.1037/0003-066x.47.11.1423

Guerin, B. (1994). Attitudes and beliefs as verbal behavior. *The Behavior Analyst*, *17*(1), 155–163. https://doi.org/10.1007/bf03392661

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Gunnarsson, K. F., Rzeszutek, M. J., Malkin, A., Tewogbola, P. O., Han, S., & Jacobs, E. A. (2026). A function-based examination of social media demand. *Behavior Analysis in Practice*. Advance online publication. https://doi.org/10.1007/s40617-026-01197-7

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Haydu, V. B., Camargo, J., & Bayer, H. (2015). Effects of preexperimental history on the formation of stimulus equivalence classes: A study with supporters of Brazilian soccer clubs. *Psychology & Neuroscience*, *8*(3), 385–396. https://doi.org/10.1037/h0101276

Hayes, S. C., Brownstein, A. J., Zettle, R. D., Rosenfarb, I., & Korn, Z. (1986). Rule-governed behavior and sensitivity to changing consequences of responding. *Journal of the Experimental Analysis of Behavior*, *45*(3), 237–256. https://doi.org/10.1901/jeab.1986.45-237

Hayes, S. C., Kohlenberg, B., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence relations. *Journal of the Experimental Analysis of Behavior*, *56*(1), 119–137. https://doi.org/10.1901/jeab.1991.56-119

Hildum, D. C., & Brown, R. W. (1956). Verbal reinforcement and interviewer bias. *The Journal of Abnormal and Social Psychology*, *53*(1), 108–111. https://doi.org/10.1037/h0040518

Houmanfar, R., & Rodrigues, N. J. (2006). The metacontingency and the behavioral contingency: Points of contact and departure. *Behavior and Social Issues*, *15*(1), 13–30. https://doi.org/10.5210/bsi.v15i1.342

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), Article e2025334119. https://doi.org/10.1073/pnas.2025334119

Insko, C. A. (1965). Verbal reinforcement of attitude. *Journal of Personality and Social Psychology*, *2*(4), 621–623. https://doi.org/10.1037/h0022485

Insko, C. A., & Melson, W. H. (1969). Verbal reinforcement of attitude in laboratory and nonlaboratory contexts. *Journal of Personality*, *37*(1), 25–40. https://doi.org/10.1111/j.1467-6494.1969.tb01729.x

Kahmann, C., & Heyer, G. (2019). Measuring context change to detect statements violating the Overton window. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 392–396. https://doi.org/10.5220/0008191803920396

Kniess, A. B., Comel, N., & Marques, F. P. J. (2024). Checagem de fatos e autoridade jornalística no Brasil: uma análise do Fato ou Fake e do Estadão Verifica [Fact-checking and journalistic authority in Brazil: An analysis of Fato ou Fake and Estadão Verifica]. *Comunicação e Sociedade*, *46*, Article e024020. https://doi.org/10.17231/comsoc.46(2024).5665

Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D., & Lewandowsky, S. (2022). Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, *1*(4), Article pgac186. https://doi.org/10.1093/pnasnexus/pgac186

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lelo, T. (2023). Assessing the consistency of fact-checking in political debates. *Journal of Communication*, *73*(6), 587–600. https://doi.org/10.1093/joc/jqad028

Levy, S., Siebold, A., Vaidya, J., Truchon, M.-M., Dettmering, J., & Mittelman, C. (2022). A look in the mirror: How the field of behavior analysis can become anti-racist. *Behavior Analysis in Practice*, *15*(4), 1112–1125. https://doi.org/10.1007/s40617-021-00630-3

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), Article 1311. https://doi.org/10.1038/s41467-020-19607-x

Linn, L. S. (1965). Verbal attitudes and overt behavior: A study of racial discrimination. *Social Forces*, *43*(3), 353–364. https://doi.org/10.2307/2574765

Lloyd, K. E. (1994). Do as I say, not as I do. *The Behavior Analyst*, *17*(1), 131–139. https://doi.org/10.1007/bf03392658

Mackinac Center for Public Policy. (2019). *The Overton window*. https://www.mackinac.org/OvertonWindow

Marino-Fages, D., & Martínez-Pozo, A. (2026). *Hate in the Tropics: Political leaders and the social acceptability of online hate speech* (Documento de trabajo N.º 391). RedNIE. https://rednie.eco.unc.edu.ar/files/DT/391.pdf

Massarani, L. M., Leal, T., Waltz, I., & Medeiros, A. (2021). Infodemia, desinformação e vacinas: a circulação de conteúdos em redes sociais antes e depois da COVID-19 [Infodemic, disinformation, and vaccines: The circulation of content on social media before and after COVID-19]. *Liinc em Revista*, *17*(1), Article e5689. https://doi.org/10.18617/liinc.v17i1.5689

Matsuda, K., Garcia, Y., Catagnus, R., & Brandt, J. A. (2020). Can behavior analysis help us understand and reduce racism? A review of the current literature. *Behavior Analysis in Practice*, *13*(2), 336–347. https://doi.org/10.1007/s40617-020-00411-4

McGlinchey, A., & Keenan, M. (1997). Stimulus equivalence and social categorization in Northern Ireland. *Behavior and Social Issues*, *7*(2), 113–128. https://doi.org/10.5210/bsi.v7i2.310

Menezes, L. F. (2024, December 23). *No segundo ano de governo, Lula desinformou mais sobre fome e educação* [In his second year in office, Lula spread the most disinformation about hunger and education]. Aos Fatos. https://www.aosfatos.org/noticias/lula-2024-desinformou-fome-educacao/

Menezes, L. F., & Faustino, M. (2025, July 7). *Como a desinformação transformou o debate político brasileiro nos últimos dez anos* [How disinformation transformed Brazilian political debate over the last ten years]. Aos Fatos. https://www.aosfatos.org/noticias/desinformacao-transformou-debate-politico-dez-anos/

Merwin, R. M., & Wilson, K. G. (2005). Preliminary findings on the effects of self-referring and evaluative stimuli on stimulus equivalence class formation. *The Psychological Record*, *55*(4), 561–575. https://doi.org/10.1007/bf03395527

Mizael, T. M. (2018). Perspectivas analítico-comportamentais sobre a homossexualidade: análise da produção científica [Behavior-analytic perspectives on homosexuality: An analysis of the scientific output]. *Revista Perspectivas*, *9*(1), 15–28. https://doi.org/10.18761/PAC.2017.011

Mizael, T. M., & de Rose, J. C. (2017). Análise do comportamento e preconceito racial: possibilidades de interpretação e desafios [Behavior analysis and racial prejudice: Possibilities for interpretation and challenges]. *Acta Comportamentalia*, *25*(3), 365–377. https://doi.org/10.32870/ac.v25i3.61632

Mizael, T. M., Coelho, C. L., Rodrigues, W. C., & de Almeida, J. H. (2021a). Racial issues and behavior analysis: Experiences and contributions from Brazil. *Behavior and Social Issues*, *30*(1), 495–513. https://doi.org/10.1007/s42822-021-00071-1

Mizael, T. M., de Almeida, J. H., Roche, B., & de Rose, J. C. (2021b). Effectiveness of different training and testing parameters on the formation and maintenance of equivalence classes: Investigating prejudiced racial attitudes. *The Psychological Record*, *71*(2), 265–277. https://doi.org/10.1007/s40732-020-00435-w

Mizael, T. M., de Almeida, J. H., Silveira, C. C., & de Rose, J. C. (2016a). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record*, *66*(3), 451–462. https://doi.org/10.1007/s40732-016-0185-0

Mizael, T. M., Dos Santos, S. L., & De Rose, J. C. C. (2016b). Contribuições do paradigma de equivalência de estímulos para o estudo das atitudes [Contributions of the stimulus equivalence paradigm to the study of attitudes]. *Interação em Psicologia*, *20*(2). https://doi.org/10.5380/psi.v20i2.46278

Moreira, M. B. (Ed.). (2025a). *Formação e mudança de atitudes: Pareamento ao modelo e história em quadrinhos* [Attitude formation and change: Matching-to-sample and comics] (1st ed.). Instituto Walden4.

Moreira, M. B. (Ed.). (2025b). *Histórias em quadrinhos e o estudo das atitudes - Parte VII* [Comics and the study of attitudes - Part VII]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_2025_hq7.pdf

Moreira, M. B., Ávila, A., Corrêa, B. F., Reis, F. G. dos, Oliveira, L. I. R. de, Pol, L. V., Silva, L. C. da, & Aguiar, S. H. A. (2023). *Histórias em quadrinhos e o estudo das atitudes - Parte II* [Comics and the study of attitudes - Part II] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2023_hq2.pdf

Moreira, M. B., Cardoso, G. P., Silveira, L. A., Carvalho, L. R., & Rabelo, M. M. (2024a). *Histórias em quadrinhos e o estudo das atitudes - Parte VI* [Comics and the study of attitudes - Part VI]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq6_2024.pdf

Moreira, M. B., Lemos, I. K., Nogueira, L., Silverio, J. P. N. A., Bacelar, J. M. C., Morais, M., Godoy, M. B. C., & da Silva, V. A. L. R. (2024b). *Histórias em quadrinhos e o estudo das atitudes - Parte III* [Comics and the study of attitudes - Part III]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq3_2024.pdf

Moreira, M. B., Martins, A. C. M., Silva, A. F. V., Ribeiro, G. M., Faria, G. M. S., & Bezerra, J. G. N. (2024c). *Histórias em quadrinhos e o estudo das atitudes - Parte IV* [Comics and the study of attitudes - Part IV]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq4_2024.pdf

Moreira, M. B., Povoa, A. M. R., Belarmino, A. J. A., Meireles, A. D. R., & Tanabe, G. Y. R. (2024d). *Histórias em quadrinhos e o estudo das atitudes - Parte V* [Comics and the study of attitudes - Part V]. Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_hq5_2024.pdf

Moreira, M. B., Premuli, I. T., Rabelo, E. de P., Pessoa, M. B. F., & Medeiros, M. R. G. C. (2022a). *Atitudes: Histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos* [Attitudes: Comics, semantic analyses of social media, and the stimulus equivalence relations paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_q0.pdf

Moreira, M. B., Rocha, G., Lago, J., Soares, A., Santos, B. dos, Simões, C., Dias, G., Nunes, J., Alves, K., Pires, M., Jardim, M., Chadud, L., Alvarenga, O., & Souza, W. (2022b). *Preconceito: Doze experimentos e um paradigma* [Prejudice: Twelve experiments and a paradigm] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_12exp.pdf

Moreira, M. B., Sousa, C. P., Tomasini, C., Craveiro, H. do A., Rodrigues, I. L., Dutra, I., Barros, J. F., Freire, L. R. V., D'arcanchy e Vilela, M. E., Cautela, M. E. T. C., Oliveira, P. M. D. N., Soares, S. K. S., & Andrade, V. M. (2022c). *Estudos sobre histórias em quadrinhos e atitudes - Parte I* [Studies on comics and attitudes - Part I] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/moreira_etal_2022_hq1.pdf

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), Article 7144. https://doi.org/10.1038/s41467-022-34769-6

Mota, A. A. S., Pimentel, S. M., & Oliveira, A. V. M. G. (2023). Desordens informativas: análise de pronunciamentos de Jair Bolsonaro contra a vacinação de covid-19 [Information disorders: An analysis of Jair Bolsonaro's public statements against covid-19 vaccination]. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, *17*(2), 311–331. https://doi.org/10.29397/reciis.v17i2.3513

Moxon, P. D., Keenan, M., & Hine, L. (1993). Gender-role stereotyping and stimulus equivalence. *The Psychological Record*, *43*, 381–394.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., . . . Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137–144. https://doi.org/10.1038/s41586-023-06297-w

Passarelli, D. A., Rico, A. S., & Silvestre, M. H. (2023). Contribuições da análise do comportamento para a compreensão do racismo [Contributions of behavior analysis to the understanding of racism]. *Revista Brasileira de Análise do Comportamento*, *19*(1). https://doi.org/10.18542/rebac.v19i1.14941

Pereira, P. B. G., & Moreira, M. B. (2021a). *Atitudes em relação a presidentes do Brasil: Formação de classes de equivalência e transferência de função* [Attitudes toward presidents of Brazil: Equivalence class formation and transfer of function] (1st ed.). Instituto Walden4. https://www.walden4.com.br/livros/pdfs/pereira_moreira_2021_presidentes.pdf

Pereira, P. B. G., & Moreira, M. B. (2021b). Transferência de atitudes em relação a presidentes da república [Transfer of attitudes toward presidents of the republic]. *Acta Comportamentalia*, *29*(4), 137–154. https://doi.org/10.32870/ac.v29i4.80318

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, *7*(6), 892–903. https://doi.org/10.1038/s41562-023-01540-w

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Recuero, R., Tavares, C., Volcan, T., Pozzebon, M., & Dutra, M. (2025). Desinformação narrada pela checagem: estudo das eleições de 2018 e 2022 [Disinformation as narrated by fact-checking: A study of the 2018 and 2022 elections]. *E-Compós*, *28*, 1–26. https://doi.org/10.30962/ecomps.3039

Regan, D. T., & Fazio, R. (1977). On the consistency between attitudes and behavior: Look to the method of attitude formation. *Journal of Experimental Social Psychology*, *13*(1), 28–45. https://doi.org/10.1016/0022-1031(77)90011-7

Ribeiro, A. (2023, December 27). *Em primeiro ano de mandato, Lula repete alegações enganosas sobre fome, economia e governo Bolsonaro* [In his first year in office, Lula repeats misleading claims about hunger, the economy, and the Bolsonaro government]. Aos Fatos. https://www.aosfatos.org/noticias/lula-alegacoes-enganosas-2023/

Ribeiro, A., & Menezes, L. F. (2022, December 22). *Uma retrospectiva das mentiras do governo Bolsonaro* [A retrospective of the Bolsonaro government's lies] [Comics feature]. Aos Fatos. https://www.aosfatos.org/noticias/retrospectiva-mentiras-governo-bolsonaro/

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-013

Risley, T. R., & Hart, B. (1968). Developing correspondence between the nonverbal and verbal behavior of preschool children. *Journal of Applied Behavior Analysis*, *1*(4), 267–281. https://doi.org/10.1901/jaba.1968.1-267

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822. https://doi.org/10.1038/s41562-023-01538-4

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Rosendo, A. P., & de Melo, R. M. (2018). Transferência de função e reorganização de classes de equivalência relacionadas a gênero e profissões [Transfer of function and reorganization of equivalence classes related to gender and professions]. *Revista Brasileira de Análise do Comportamento*, *14*(1), 31–43. https://doi.org/10.18542/rebac.v14i1.7157

Santos, L. F. dos. (2022, August 17). *Bolsonaro supera 6.000 declarações falsas ou distorcidas no início da campanha pela reeleição* [Bolsonaro surpasses 6,000 false or distorted statements at the start of his reelection campaign]. Aos Fatos. https://www.aosfatos.org/noticias/bolsonaro-supera-6000-declaracoes-falsas-ou-distorcidas-no-inicio-da-campanha-pela-reeleicao/

Scott, W. A. (1957). Attitude change through reward of verbal behavior. *The Journal of Abnormal and Social Psychology*, *55*(1), 72–75. https://doi.org/10.1037/h0040249

Shea, P., Johnson, P., & Togade, D. (2023). Using relational frame theory to examine racial prejudice: A tool for educators and an appeal for future research. *Behavior Analysis in Practice*, *16*(1), 102–116. https://doi.org/10.1007/s40617-022-00767-9

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. Matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 U.S. election campaign and its aftermath. *Quarterly Journal of Political Science*, *16*(1), 71–104. https://doi.org/10.1561/100.00019045

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Spielberger, C. D., Levin, S. M., & Shepard, M. G. (1962). The effects of awareness and attitude toward the reinforcement on the operant conditioning of verbal behavior. *Journal of Personality*, *30*(1), 106–121. https://doi.org/10.1111/j.1467-6494.1962.tb02308.x

Street, W. R. (1994). Attitude-behavior congruity, mindfulness, and self-focused attention: A behavior-analytic reconstruction. *The Behavior Analyst*, *17*(1), 145–153. https://doi.org/10.1007/bf03392660

Tsipursky, G., Votta, F., & Roose, K. M. (2018). Fighting fake news and post-truth politics with behavioral science: The Pro-Truth Pledge. *Behavior and Social Issues*, *27*(1), 47–70. https://doi.org/10.5210/bsi.v27i0.9127

Visser, M. (1996). Voting: A behavioral analysis. *Behavior and Social Issues*, *6*(1), 23–34. https://doi.org/10.5210/bsi.v6i1.278

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Weinstein, J. H., Wilson, K. G., Drake, C. E., & Kellum, K. K. (2008). A relational frame theory contribution to social categorization. *Behavior and Social Issues*, *17*(1), 40–65. https://doi.org/10.5210/bsi.v17i1.406

Zielinski, C. (2021). Infodemics and infodemiology: A short history, a long future. *Revista Panamericana de Salud Pública*, *45*, 1. https://doi.org/10.26633/rpsp.2021.40

APPENDIX

D

Source Verification

PROVENANCE NOTE

The verification recorded in this appendix was carried out by an artificial intelligence (Claude, by Anthropic), which assisted in writing the book, and not by the authors. It was the AI that assembled the spreadsheet of claims, obtained or received the texts of the sources through legal means, searched them for the sentences that support each claim, read the context, and recorded the results and the caveats.

For this reason, "read" and "not read," in Sections D.2 to D.4, state what this review was able to read and check, not what the authors read or did not read. A "not read" source is a source that the review could not verify: the book does not treat it as verified and, where the subject required the source, the chapter states that it was not read.

This appendix documents how the claims in the chapters were checked against the sources, how many sources were read and which were left out, and which methodological decisions affect the reading.

D.1 How a claim comes to be "verified"

Each claim attributed to a source is entered in a spreadsheet with an identifier, the source key, the chapter, the statement, and the verbatim sentences from the source that support it. A script searches for these sentences in the extracted full text (ignoring spaces, accents, punctuation, hyphenation, and ligatures), records the page, and returns the context; the decision rests on reading the context, which is recorded along with the date and the caveats. Only then does the claim become "verified" and eligible to be cited. The text of each chapter marks, in comments, the claims that support each paragraph, and a second script checks that every in-text citation has a reference and that every marked claim exists and is verified.

D.2 Numbers

The survey records 168 sources, of which 108 were obtained through legal means and read in full text and 60 remain pending (Section D.4). The book cites 124 works. The claims spreadsheet has 1,207 entries, all recorded as verified, with the context read. The complete spreadsheet (reference, DOI, excerpt and page, status) is in verificacao/tabela-de-verificacao.xlsx.

D.3 Versions read that are not the journal version

When the version read is not the published one, or does not carry the journal's pagination, the chapter cites without a page number and says so. The sources cited in the book in this situation:

- Ajzenman et al. (2023): submitted version, on RePEc.
- Álvarez-Benjumea and Winter (2018): submitted version, on MPG.PuRe.
- Bail et al. (2018): XML text (PMC), without the journal's pagination.
- Brady et al. (2021): XML text (PMC), without the journal's pagination.
- Calleja et al. (2021): XML text (PMC), without the journal's pagination.
- Ceylan et al. (2023): XML text (PMC), without the journal's pagination.
- Cinelli et al. (2021): XML text (PMC), without the journal's pagination.
- Effron and Raj (2020): accepted manuscript, on London Business School Research Online.
- Eysenbach (2009): XML text (PMC), without the journal's pagination.
- Eysenbach (2020): XML text (PMC), without the journal's pagination.
- Ferster and Skinner (1957): original 1957 edition, Appleton-Century-Crofts, PsycBooks copy from the author's collection.
- Frimer et al. (2023): accepted manuscript, on PsyArXiv.
- Garrett and Bond (2021): XML text (PMC), without the journal's pagination.
- Guess et al. (2019): XML text (PMC), without the journal's pagination.
- Guess et al. (2020): XML text (PMC), without the journal's pagination.
- Hangartner et al. (2021): XML text (PMC), without the journal's pagination.
- Huszár et al. (2022): XML text (PMC), without the journal's pagination.
- Lasser et al. (2022): XML text (PMC), without the journal's pagination.
- Lazer et al. (2018): submitted version, on arXiv.

- Mackinac Center (2019): web page, an interested source.
- Marino-Fages and Martínez-Pozo (2026): working paper, not peer reviewed.
- Mota et al. (2023): submitted version, on LA Referencia.
- Moxon et al. (1993): ResearchGate copy (scan of the published PDF).
- Nyhan and Reifler (2010): author's manuscript, without the journal's pagination.
- Rathje et al. (2021): submitted version, on Apollo.
- Roozenbeek et al. (2022): submitted version, on Apollo.
- Shea et al. (2023): XML text (PMC), without the journal's pagination.
- Siegel et al. (2021): version posted by the author.
- Skinner (1957): original 1957 edition, Appleton-Century-Crofts, copy from the author's collection.
- Zielinski (2021): XML text (PMC), without the journal's pagination.

D.4 Sources not read

Sources planned in the survey for which the full text could not be obtained through legal means. They are not cited as verified; where the subject would require the source, the chapter states that it was not read. Of the 60 pending sources in the survey, 24 have an attempt to obtain them through open access (repository, PMC, author's version) recorded in verificacao/fontes.tsv; the others had no open access located, and both groups depend on the author attaching the PDF. The last group in the list consists of the author's own books, which fall outside the exception in the source rules (Section D.5).

Open access, but the automatic download was blocked (24)

- Biglan, A., & Embry, D. D. (2013). A framework for intentional cultural change. *Journal of Contextual Behavioral Science*, *2*(3-4), 95–104 (Chapter 12).
- Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*(1), 127–146 (Chapter 4).
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the implicit relational assessment procedure (IRAP) and the relational elaboration and coherence (REC) model. *The Psychological Record*, *60*(3), 527–542 (Chapter 4).

- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29 (Chapters 1, 12).
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002 (Chapter 3).
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880 (Chapter 3).
- Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, *5*(3), Article 2053168018786848 (Chapter 2).
- Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3–33 (Chapter 5).
- Ozawa, J. V. S., Woolley, S. C., Straubhaar, J., Riedl, M. J., Joseff, K., & Gursky, J. (2023). How disinformation on WhatsApp went from campaign weapon to governmental propaganda in Brazil. *Social Media + Society*, *9*(1), Article 20563051231160632 (Chapter 6).
- Rossini, P., Stromer-Galley, J., Baptista, E. A., & Veiga de Oliveira, V. (2021). Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society*, *23*(8), 2430–2451 (Chapter 6).
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318 (Chapter 8).
- Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science*, *27*(7), 1027–1035 (Chapter 9).
- Juul, J. L., & Ugander, J. (2021). Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, *118*(46), Article e2100786118 (Chapter 11).
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559 (Chapter 11).

- Galhardi, C. P., Freire, N. P., Minayo, M. C. S., & Fagundes, M. C. M. (2020). Fato ou Fake? Uma análise da desinformação frente à pandemia da Covid-19 no Brasil [Fact or fake? An analysis of disinformation in the face of the Covid-19 pandemic in Brazil]. *Ciência & Saúde Coletiva*, *25*(Suppl. 2), 4201–4210 (Chapter 11).
- Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1), Article 20563051221150412 (Chapter 1).
- Chadwick, A., Vaccari, C., & Kaiser, J. (2025). The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology. *American Behavioral Scientist*, *69*(2), 113–130 (Chapters 8, 9).
- Bastos dos Santos, J. G., Freitas, M., Aldé, A., Santos, K., & Cardozo Cunha, V. C. (2019). WhatsApp, política mobile e desinformação: a hidra nas eleições presidenciais de 2018 [WhatsApp, mobile politics, and disinformation: The hydra in the 2018 presidential elections]. *Comunicação & Sociedade*, *41*(2), 307–334 (Chapter 6).
- Mercuri, K. T., & Lima-Lopes, R. E. (2020). Discurso de ódio em mídias sociais como estratégia de persuasão popular [Hate speech on social media as a strategy of popular persuasion]. *Trabalhos em Linguística Aplicada*, *59*(2), 1216–1238 (Chapter 5).
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402 (Chapters 1, 3).
- Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30 (Chapters 1, 5).
- Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), Article e2104235118 (Chapter 11).
- Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67 (Chapter 7).
- Cazzamatta, R., & Santos, A. (2024). Checking verifications during the 2022 Brazilian run-off election: How fact-checking organizations exposed falsehoods and contributed to the accuracy of the public debate. *Journalism*, *25*(10), 2022–2043 (Chapter 2).

No open access located (36)

- Verplanck, W. S. (1955). The control of the content of conversation: Reinforcement of statements of opinion. *The Journal of Abnormal and Social Psychology*, *51*(3), 668–676 (Chapters 3, 9).
- Bem, D. J. (1965). An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology*, *1*(3), 199–218 (Chapter 3).
- Silver, B. D., Abramson, P. R., & Anderson, B. A. (1986). The presence of others and overreporting of voting in American national elections. *Public Opinion Quarterly*, *50*(2), 228–239 (Chapters 1, 7).
- Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, *52*(5), 270–277 (Chapter 3).
- Skinner, B. F. (1981). Selection by consequences. *Science*, *213*(4507), 501–504 (Chapters 6, 9, 11).
- Glenn, S. S., Malott, M. E., Andery, M. A. P. A., Benvenuti, M., Houmanfar, R. A., Sandaker, I., Todorov, J. C., Tourinho, E. Z., & Vasconcelos, L. A. (2016). Toward consistent terminology in a behaviorist approach to cultural analysis. *Behavior and Social Issues*, *25*(1), 11–27 (Chapters 6, 9, 11).
- Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research*, *14*(1), 5–13 (Chapter 4).
- Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record*, *41*(1), 33–50 (Chapter 4).
- Grey, I. M., & Barnes, D. (1996). Stimulus equivalence and attitudes. *The Psychological Record*, *46*, 243 (Chapter 4).
- Dixon, M. R., Rehfeldt, R. A., Zlomke, K. R., & Robinson, A. (2006). Exploring the development and dismantling of equivalence classes involving terrorist stimuli. *The Psychological Record*, *56*(1), 83–103 (Chapter 4).
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378 (Chapter 1).
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50 (Chapter 1).
- Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113 (Chapter 1).

- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112 (Chapter 3).
- Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, *25*(2), 162–180 (Chapter 2).
- Marietta, M., Barker, D. C., & Bowser, T. (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, *13*(4), 577–596 (Chapter 2).
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375 (Chapters 2, 12).
- Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146 (Chapter 5).
- Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167 (Chapter 5).
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132 (Chapter 6).
- González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M., Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., . . . Tucker, J. A. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, *381*(6656), 392–398 (Chapter 6).
- Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*(1), 33–48 (Chapter 7).
- Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771 (Chapter 8).
- Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., . . . Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404 (Chapter 8).
- Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine*, *113*(9), 763–765 (Chapter 11).

- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131 (Chapter 12).
- Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441 (Chapter 12).
- Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163 (Chapter 12).
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595 (Chapter 12).
- Lelo, T. (2023). Assessing the consistency of fact-checking in political debates. *Journal of Communication*, *73*(6), 587–600 (Chapter 2).
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153 (Chapter 1).
- Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication*, *37*(2), 145–156 (Chapter 1).
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257 (Chapter 3).
- Batista Pereira, F., Bueno, N. S., Nunes, F., & Pavão, N. (2022). Fake news, fact checking, and partisanship: The resilience of rumors in the 2018 Brazilian elections. *The Journal of Politics*, *84*(4), 2188–2201 (Chapters 6, 11).
- Müller, K., & Schwarz, C. (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, *15*(3), 270–312 (Chapter 5).
- Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting. *Public Opinion Quarterly*, *65*(1), 22–44 (Chapter 7).

Books by the author outside the exception (3)

- Efeitos reforçadores da atenção sobre o comportamento humano [Reinforcing effects of attention on human behavior] (book by the author, outside the exception in the source rules) (Chapters 8, 9).
- Comportamento e práticas culturais [Behavior and cultural practices] (book by the author, outside the exception in the source rules) (Chapters 9, 10).

- Modificação do comportamento verbal através de reforçamento em intervalo variável em situação análoga à psicoterapia: uma replicação sistemática [Modification of verbal behavior through variable-interval reinforcement in a situation analogous to psychotherapy: A systematic replication] (book by the author, outside the exception in the source rules) (Chapter 9).

D.5 Methodological decisions that affect the reading

News reports are not used as a source of data. The project's rule for news reports (from the Brazilian outlets G1, CNN Brasil, BandNews, and Canal Meio, with a screenshot, URL, and dates of publication and access) requires collecting pages and images, which could not be done in an auditable way; the cases of statements are those documented in articles and reports whose full text was read, cited with the page. The rule remains in force for anyone who adds news reports.

The statements in Chapter 3 are illustrative cases, not a sample. The comparable sample in Chapter 2 (Section 2.9) has a fixed rule and was not collected: Lupa and Aos Fatos prohibit artificial intelligence crawlers in their robots.txt files, the public ClaimReview feed covers Lupa very unevenly, and Lupa's report does not break down Lula's terms in office. Before any request to a website, its robots.txt was read in a separate step.

Two books are included by exception, approved by the author: Skinner (1957, *Verbal Behavior*) and Ferster and Skinner (1957, *Schedules of Reinforcement*), read in copies from the author's collection, with the printed page numbers. Other books are excluded, except the author's books on attitudes and on comics. Three other books by the author, on the reinforcing effects of attention, on cultural practices, and a systematic replication on the modification of verbal behavior, are not covered by this exception: they were not read and are not cited, and they would strengthen Chapters 8 to 10 if the author extends the exception. The author's manual "O conceito de atitude na Análise do Comportamento" [The concept of attitude in behavior analysis] remains unpublished and is not cited; the layered definition of attitude (Chapter 4, Section 4.1) is anchored in published works.

The cropped images from the Lupa report from which Chapter 3 reads the classifications are in verificacao/evidencias/, outside the text of the book.

D.6 Recorded procedural slips

Two methodological slips are recorded because the book states that the robots.txt was read before any request. On October 11, 2026, to obtain metadata for Frimer et al. (2023), a request was sent to api.osf.io before checking the robots.txt, which on that site reads Disallow: /; only abstract metadata came back, which were not used, and the text read is that of the PDF obtained from osf.io/download. On October 14, 2026, while trying to obtain the text of two sources through PMC, a request to the efetch service of eutils.ncbi.nlm.nih.gov was made in the same command that printed the robots.txt, also Disallow: /; only metadata without body text came back, nothing was used, and the files were deleted. Since then the robots.txt has been read in a separate step, and the NCBI E-utilities are not used.

PUBLISHER

Instituto Walden4 Press

Instituto Walden4 Press is dedicated to promoting and disseminating advanced knowledge in behavior analysis, as well as in other evidence-based applied fields, covering both scientific and professional aspects. Committed to democratizing access to information, it offers a wide range of books free of charge, and all titles are available in digital format.

EDITORIAL BOARD

Dr. Gleidson Gabriel da Cruz

Dr. Márcio Borges Moreira

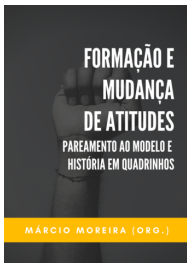
CONTACT

secretaria@walden4.com.br

walden4.com.br · @instituto.walden4

CATALOG

Other books from Instituto Walden4 Press



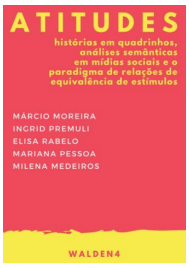
Formação e Mudança de Atitudes: Pareamento ao Modelo e História em Quadrinhos

Márcio Borges Moreira (Ed.) · 2025



Preconceito: doze experimentos e um paradigma

Márcio Borges Moreira · 2022



Atitudes: histórias em quadrinhos, análises semânticas em mídias sociais e o paradigma de relações de equivalência de estímulos

Márcio Borges Moreira · 2022



Atitudes em relação a presidentes do Brasil: formação de classes de equivalência e transferência de função

Pedro Bufara Gonzaga Pereira, Márcio Borges Moreira · 2021



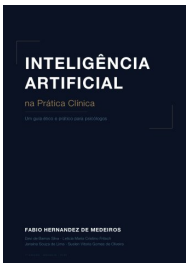
Efeitos reforçadores da atenção sobre o comportamento humano

Márcio Borges Moreira · 2021



Comportamento e Práticas Culturais

2015



Inteligência Artificial na Prática Clínica

Medeiros et al. · 2026



Do Encanto às Contingências

2026



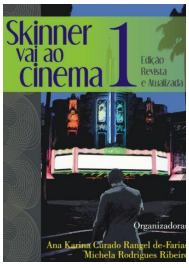
Uma história de aprendizagem operante

Márcio Borges Moreira, Lucas Couto de Carvalho · 2017



Experimentos clássicos em Análise do Comportamento

2016



Skinner Vai ao Cinema (Vol. 1)

2015



A Psicologia como Estudo de Interações

João Claudio Todorov · 2015

See the full catalog (in Portuguese), with 65 titles, many of them free

walden4.com.br/livros