PAIRING AFTER THE FALL

PAIRING
after the
FALL

How men and women could live together,
from the couple to the planet

Book Two of *THRIVING after the FALL*

THE DATA EDITION

Directed and produced by Thomas Schindler

Written by an LLM with the support of humanity

Copyright © 2026 Thomas Schindler.
Directed and produced by Thomas Schindler. Written by an LLM with the support of humanity.

All rights reserved.

The data edition. First draft of 1 October 2026, with every figure and its source in the text. Its figures are still being checked against the original papers.

The studies, figures, sources and open questions behind every chapter are at thriving.thomas.cr/pairing/

Contents

Preface vii

Part One Overlapping Bells 1

1. The Grandmother Clause 3
2. The Campfire Contract 19
3. The Endocrine Accent 35
4. Overlapping Bells 51
5. Two Shelves and a Trap 69

Part Two The Campfire Contract 81

6. Flooding 83
7. The Second Shift That Wasn't 99
8. The Father's Falling Tide 115
9. Who Keeps Children Alive 133
10. The Factor That Fell 151

Part Three Emancipation 2.0 167

11. The Abbess and the Monks 169

12. The Bus Route 185

13. The Wall 201

14. Queens at War 217

15. The Long Argument 233

16. Emancipation 2.0 249

Preface

The other half of the question

On the sixth of June 1971 the West German magazine *Stern* went on sale with 374 women's faces on its cover and three words printed across them: *Wir haben abgetrieben!* We have had abortions. Romy Schneider was in the grid. So were secretaries, students and housewives nobody had heard of, each confessing in print to a crime under Paragraph 218 of the criminal code. Alice Schwarzer had organised it, borrowing the idea from a French manifesto of 343 women published that spring. It was the loudest single sentence West German women had spoken since the war, and it was a sentence about their own bodies.

My mother was twenty-five that summer. Her name was Barbara. She was born in 1946, in the first year of the peace, in Lobberich, a small town on the Lower Rhine a few kilometres from the Dutch border, and she died in 2021, fifty years after that cover. My father was three years older, born in 1943. He died in 2001. I arrived in 1975, which puts me in the middle of the argument this book is about. What I can tell you is the country she was living in. A married woman there could not take a job against her husband's wishes until the marriage law was reformed in 1977. The tax code had, since 1958, pooled a couple's income and split it in two, which sounds fair and in practice meant that her first Mark of wages was taxed at his top rate. Schools sent the children home at lunchtime. Public crèches for children under three were, for practical purposes, a rumour. And for the mother who worked anyway, the language had a word waiting: *Rabenmutter*. Raven mother. The bird that leaves the nest. For the father who left it every morning at seven, the language had nothing at all.

Against all of that, her generation fought for one simple, radical idea: a

woman can do anything a man can do. They were right. They won more than they lost. And they were, it turns out, answering half of a question.

The other half, *and should she want to do it the same way, and should he?*, got filed under reactionary and left in a drawer for fifty years.

This book opens the drawer.

* * *

The other side of the wall

On the far side of the border, with the same language, the same grandmothers and the same Sunday cake, the state had taken the opposite bet. The constitution of the German Democratic Republic promised women the right to work and equal pay from 1949. By the late 1980s public childcare covered nearly every child in the country. From 1976 a mother got the *Babyjahr*, a paid year off with her job held open. The insult ran the other way, too. A woman who stayed home with the children was a *Schmarotzer*, a parasite on the collective. In 1989, the last full year before the wall came down, 89 per cent of East German women of working age were in the labour force. In the West it was 56.

Then the West's laws, taxes and half-day schools were rolled out across the East, and the experiment should have ended.

It did not. Three economists, Barbara Boelmann, Anna Raute and Uta Schönberg, went through the social-security records of half the German workforce and found that, a full generation after reunification, mothers in the East still work about 350 hours a year more than mothers in the West. Fathers: no gap at all. Seven years after a first child an East German mother is back at 70 per cent of her old earnings, a West German mother at 45. Wages and the price of childcare explain none of it. Their estimate is that the working-mother stigma alone accounts for 73 per cent of the gap.

It goes further. In the West, when a wife's income crosses half the household total, something odd happens in the following year. She does more housework, not less. She is more likely to leave her job. The couple is more likely to divorce. Sociologists call it "doing gender", the quiet restoration of a hierarchy that the payslip has just upset. In the East, none of it happens. East German women out-earn their husbands and nobody reaches for the ironing board or the lawyer. That is twenty-one years of panel data, 1991 to 2012, analysed by Quentin Lippmann, Anne Georgieff

and Claudia Senik in *The Economic Journal*. And when West German mothers move east, within a few years they work like East German mothers. When East German mothers move west, they keep working like East German mothers, tax code, stigma and all.

If you wanted one experiment to show that culture, not hormones, decides who stays home with the baby, you could not design a better one. Nobody designed it. It happened to my parents' generation, with a wall down the middle.

So why isn't this a short book?

* * *

The body does not read the newspaper

Because the body was never consulted, and it kept measuring anyway.

Put a hand-grip dynamometer in front of 2,187 people, as Dieter Leyk and his colleagues did in 2007, and the men squeeze an average of 541 newtons and the women 329. Ninety per cent of the women fall below 95 per cent of the men. Women who compete in sport at national level land at about the 25th percentile of untrained men. In the statistician's language that is a gap of more than two standard deviations, and two standard deviations is not a "tendency". It is a different piece of equipment. Men carry roughly 60 per cent more muscle than women, and most of the extra is above the waist (William Lassek and Steven Gaulin, 2009).

Now the other column, which the strong-man reading always leaves out. Women outlive men in every famine and epidemic for which the records survive (Virginia Zarulli and colleagues, *PNAS*, 2018). Neolithic farming women in Central Europe, ground down at the quern stone for hours a day, had upper-arm bones 11 to 16 per cent stronger than those of today's elite rowers (Alison Macintosh, 2017). Women carry about 1.6 times the body fat of men, and the fat on the hips and thighs is a store of exactly the fatty acids an infant's brain is built from. A woman pays the whole metabolic bill for gestation and lactation, which no tax reform has yet found a way to split. And at around fifty her ovaries stop, decades before the rest of her does, which almost no other mammal on earth has thought worth doing. Hold that thought. It may be the most important sentence in the book.

Then the hormones, which is where my mother's generation stopped listening and where this book begins. Testosterone does not make men ag-

gressive in any simple way. The correlation between a man's baseline level and his aggression is close to zero across 45 studies (Shawn Geniole, 2020). What testosterone does is *respond*. It rises with competition and with winning, in both sexes. It falls, sharply and lastingly, when a man becomes a father who actually holds the baby (Lee Gettler's cohort of young men in Cebu, in the Philippines, *PNAS*, 2011), and prolactin and oxytocin rise to meet it. The most beautiful experiment in the field gave a group of women in Zurich a dose of testosterone under the tongue and sat them down to a bargaining game. The hormone made their offers *fairer*. The women who merely believed they had been given testosterone made meaner ones (Christoph Eisenegger and colleagues, *Nature*, 2010). The molecule did one thing. The folklore about the molecule did the opposite.

That is the pattern this book keeps finding, and it needs a name. I will call it the *Endocrine Accent*. Hormones do not write the script. Everyone in the room speaks the same language of fear, ambition, tenderness, boredom and lust. Not everyone speaks it with the same lilt, and the lilt is measurable, and on average it is not the same in the two sexes, and it shifts with what life throws at you. The trick, for a couple or a company or a country, is to hear the accent without mistaking it for the message.

* * *

Overlapping bells

"On average" did a lot of work in that paragraph, so let me pay for it now.

In 2005 Janet Hyde pooled 46 meta-analyses of psychological sex differences and found that 78 per cent of them were small or near zero. Ten years later Ethan Zell and colleagues did it again with 106 meta-analyses and more than twelve million people: the average difference came to 0.21 of a standard deviation. Draw two bell curves that far apart and they overlap by about 85 per cent. Pick a random woman and a random man, test them on some "male-typical" trait, and she outscores him 42 times in a hundred. On maths, across 1.3 million pupils, the gap is 0.05 (Sara Lindberg, 2010). On verbal ability it tilts slightly the other way. Those are the *Overlapping Bells*, and any sentence in this book that begins "men are" or "women are" is to be read with the two bells in mind. An average is a fact about a population. It is never a rule for the person across the table.

But the bells are not the whole story either, and a fair book has to say

so. Some differences are large. Interest in things against interest in people: 0.93 of a standard deviation across half a million people (Rong Su, James Rounds and Patrick Armstrong, 2009), stable across decades and continents, and the single best-supported explanation for why engineering faculties and nursing schools look the way they do, even in Norway. Physical aggression: about 0.45. Rotating a shape in the mind's eye: about 0.57. And in 2012 Marco Del Giudice and colleagues asked a different question. Not "how far apart are men and women on trait X?" but "if you look at the whole personality at once, how far apart are the two clouds?" The answer, on 10,261 Americans, was 2.71 standard deviations, an overlap of about 10 per cent. A replication in 2019 got 2.06 and about 30 per cent. Critics say the method piles up a heap of small differences into one big one, and they have a point, and the argument is not over. But the shape of the finding is the shape of this book's central puzzle: *similar trait by trait, distinct in combination.* Which is also, if you think about it, how you recognise a face.

The brain follows the same rule. Correct for the fact that male brains are about 10 per cent bigger and sex explains about 1 per cent of the remaining variation in structure (Lise Eliot, 2021, in a paper titled "Dump the dimorphism"). Look at any one brain and it is a mosaic of "male-typical" and "female-typical" features, only a few per cent of brains being consistent all the way through (Daphna Joel, *PNAS*, 2015, more than 1,400 scans). Feed the same kind of scans to a classifier and it guesses sex correctly 93 times in 100 (Adam Chekroud, *PNAS*, 2016). Both are true. They are not a contradiction. They are the subject.

* * *

Two shelves and a trap

Walk into a bookshop and the shelf on men and women is really two shelves facing each other across an aisle nobody crosses. On one side the evolutionary architects: Pleistocene mating markets, hunters and gatherers, the sexes as two strategies. On the other the constructivists: the mosaic brain, the gendered toy aisle, difference as something done to us. Each shelf has its own data and its own villains. Each is, in places, right. And each is written, mostly, as a weapon. The reader gets to choose a side and go home confirmed, which is a service of a kind. It is not the service this book is for.

The trap has a philosophical name and two jaws. The first jaw is the naturalistic fallacy: because something is natural, it must be right. David Hume spotted it in 1739. G. E. Moore gave it its name in 1903. Lethal male violence has deep evolutionary roots. So does infanticide. Nobody sane concludes we ought to keep them. The second jaw is newer and less often named: the moralistic fallacy, christened by the microbiologist Bernard Davis in *Nature* in 1978. Because something would be unfair if it were true, it cannot be true. This one is my mother's generation's tooth, and I say that with love. If women and men differed on average in what they wanted from a life, then fifty years of policy built on interchangeability might need a rethink, and that is unthinkable, therefore they don't differ. I have heard the argument at dinner. I have made it.

This book will try to keep both jaws open. It will fail here and there. When it does, the notes at the back will let you catch it.

* * *

What fell

One more thing before I tell you the plan, because you have been sold this subject before and some of what you were sold has since been recalled.

Power posing, the two minutes of standing like Wonder Woman that were supposed to raise your testosterone and lower your cortisol before the interview: a pre-registered replication with a far larger sample found no effect on either hormone or on risk-taking, and the lead author of the original study has publicly disowned it, which is more than most findings get. The finding that women prefer square-jawed men on their fertile days: gone, once researchers measured hormones instead of counting days on a calendar. Single-sex schooling as an academic advantage: 184 studies, 1.6 million pupils, 21 countries, and once you allow for who gets sent to those schools the benefit is a rounding error (Erin Pahlke, Janet Hyde and Carlie Allison, 2014). Women talk more than men: recorders clipped to the collar say they don't. Oxytocin, the "cuddle hormone" that would make us all trust each other: the big meta-analyses find no such thing, though it does reliably make us warmer to our own side and colder to the other, which is a finding, just not a cuddly one. And for the other shelf, the belief that women in power are the peaceful sex: across 193 European reigns between 1480 and 1913, queens were 39 percentage points *more* likely than kings to be at war in any given year (Oeindrila Dube and S. P. Harish, 2020).

Then there is the flagship of my own side of the argument. Sex differences in personality and in career choice get *larger*, not smaller, in the richest and most equal countries: Sweden, Finland, Norway. Real as a pattern, across tens of thousands of people in dozens of countries. Fiercely contested as an explanation. The best counter-study finds that the stereotype "maths is not for girls" is *stronger* in rich countries, and once you allow for it the paradox goes away (Thomas Breda and colleagues, *PNAS*, 2020, 300,000 pupils). I don't know who is right. Neither, honestly, does anyone else yet. You will get both.

Every chapter of this book ends with a short list: what in it is load-bearing, what is decoration, and what is still being fought over.

* * *

How the book goes about it

The method, in four rules.

One. Three layers, never mixed. Physiology is *measurable*: muscle, fat, hormones, the ovary that stops. Behaviour is *statistical*: two overlapping bells, real at the level of a population, silent at the level of a person. Culture is *variable*: a wall down the middle of a country moved a bell curve by 350 hours a year. Almost every bad argument about men and women, from either shelf, is a claim from one layer smuggled into another. Whenever this book says a thing is true, it will say which layer it is true in.

Two. Conditional sentences only. Not "men are X, therefore society must do Y". Instead: *given what we are, arrangement X under conditions Y tends to produce Z.* It is a less thrilling grammar. It is the only one that survives contact with the German border.

Three. The strongest case against, every time. Each chapter carries a passage, set apart, that gives the best evidence against what the chapter has just argued, in that evidence's own numbers and its own best voice. Not a straw man. The real thing. If the chapter cannot survive the passage, the chapter gets rewritten, not the passage.

Four. Numbers with names on them. Every figure in this book stands on a named study that you can go and find. Where a figure could not yet be checked against the original paper, the text says so.

And the route. The book has three parts.

Part I. Overlapping Bells. The difference, as far as it can be measured. It starts not with us but with the animals that solved this before we

did: the killer whale pods with a grandmother at the front, the bonobo sisterhoods that outrank bigger males by ganging up, the elephant bulls who calm the teenagers down. Then the deep history: the campfire economics of foraging bands, which divided work and power without writing anything down and, on the whole, without the wall. I will call it the *Campfire Contract*. Then the body and its accent. Then the mind, the bells, and the argument about combinations.

Part II. The Campfire Contract. What the difference does at the scales where most of life is lived. The couple, where the best-known predictor of divorce turns out to be a risk factor, not a prophecy. The family, where a father's testosterone drops and a grandmother's presence changes whether a child thrives. The small group, where the share of women in a team predicts how well it thinks, or does not, depending on which study you read.

Part III. Emancipation 2.0. The organisation, the city, the state, the world. What the German experiment, the kibbutz, the Shakers and the Oneida colony actually taught. Why Vienna redesigned its parks and its bus routes. Why the best single predictor of whether a state is peaceful, across 175 countries, is how it treats its women, and why that is not the same as saying women are peaceful. And finally the question my mother's generation asked, asked again with the second half attached: not only *can* she, but *how would she and he want to*, and what would a world built for that look like.

Seven scales, from the couple to the planet. At each one the same three questions: what is measurable, what is statistical, and what is ours to decide.

* * *

How this book was made

I should say how it was researched, because it bears on how far you should trust it. The library behind this book is 81 reports of 15 to 20 pages each, written by machine: deep-research models working from briefs I wrote, each brief asking for the case against as well as the case for. The prose you are reading was drafted by a machine to rules I set and in a voice I chose, and then argued over. Every figure in it is being checked by hand against the original paper, and where that check has not yet happened the text says so. I mention this not as an apology but as a method. A ma-

chine has no mother, no generation and no side. It also has no judgement, which is where I come in, and where you do.

* * *

THE STRONGEST CASE AGAINST THIS BOOK

Rule three applies to prefaces too.

The strongest case against this book is that the wall wins. That a tax code, a school timetable and one well-aimed insult can move the whole behaviour of a population's mothers, while the hormones sit in the background explaining 1 per cent of brain structure and a fifth of a standard deviation of personality. On that reading my mother's generation missed nothing that matters for how we live, and the endocrine accent is real but irrelevant: a fact about bodies with no lesson for institutions, and any book that goes looking for one is the naturalistic fallacy in a nicer jacket.

I take that seriously. I have spent more of the past year with that argument than with my own. Here is why I still wrote the book.

The wall moved the mothers. It did not move the fathers, in either direction. The East abolished the breadwinner norm and kept the same woman doing most of the laundry, most of the cooking and most of the thinking about both: in reunified Germany's family panel, women are still the one primarily responsible for laundry in 71 per cent of couples and for cooking in 58, and couples who start out symmetrical revert to an unequal split at the birth of the first child, almost invariably, East and West alike. About one woman in five in such a household is thinking about leaving. And in the same panel, men who cut their paid hours and take on more of the routine domestic work report *higher* life satisfaction, not lower. Something here is not being explained by the wall. It might be a second, subtler wall. It might be the accent. The honest answer is that we do not know yet, and a book that assumed either answer would not be worth your evening.

* * *

The grandmother in front

Somewhere off the Pacific coast of North America a pod of killer whales is travelling in a line, and the animal at the front is a female who has not had a calf in twenty years.

Menopause has been found in exactly five species of toothed whale and in us. In 2024 a team led by Samuel Ellis compared 18 whale species in *Nature* and found that the ones with menopause did not shorten their reproductive lives. They lengthened everything else, by about forty years. The body keeps going long after it has stopped making the next child, and it keeps going for a reason. In the Southern Resident orcas, when a post-reproductive mother dies, her adult sons' risk of dying in the following year rises many times over, by one analysis nearly fourteen-fold (Emma Foster and colleagues, *Science*, 2012). When a grandmother dies, her grandcalves' mortality rises four and a half times (Stuart Nattrass and colleagues, *PNAS*, 2019). She is not decoration. She is infrastructure.

Nobody has yet satisfactorily explained why a body would switch off its own reproduction and go on living for three decades, unless what that body is *for*, in that species, is something other than making the next child. I will call this the *Grandmother Clause*, and it is where the book begins: with an animal that has never heard of Paragraph 218 and has nonetheless organised its whole society around the difference between its sexes, in a way that puts an old female in front.

My mother's generation answered half the question. This book is about the other half.

It begins in the water.

PART ONE

Overlapping Bells

CHAPTER 1

The Grandmother Clause

Hold your breath, and stay under for a moment.

Somewhere in the cold green water of the North Pacific, a line of killer whales is moving in one direction, close enough to touch, and the animal at the front is old. She is a female. Her last calf is far behind her, possibly decades behind her, and the whales at her back include daughters with calves of their own and at least one grown son who has never left her. Nobody voted. Nobody had to. The pod goes where she goes, and in a year like this one, when the Chinook salmon are thin, it goes there more than usual.

A little way back in the line, a mother takes a salmon crossways in her jaws and bites it in two. Half of it drifts to the male beside her. He is her son. He is fully grown and perfectly able to catch his own. The people who watch this population have her down as K16, and they have watched her do it for years.

That is the whole chapter in one picture. Here is a society organised, top to bottom, around a real difference between its sexes: one sex stops breeding at about forty and lives to eighty or ninety, and the other does not. And the society has answered that difference by putting the female who has stopped breeding at the front and feeding the male who has not. The order of the line is the point. Who swims first, and why, and whether it changes when the fish run out.

I want to stay in the water for a while, then climb out and walk through an elephant herd, then sit with two apes who are equally our cousins and disagree about almost everything, and then take a detour through a shore-

bird colony. The question underneath all of them is simpler than any one animal and harder than all of them together. When somebody says that an arrangement between men and women is *natural*, what do they think nature does? The honest answer, as you will see, is that nature does everything, in every direction, sometimes in the same species on the same afternoon. That is not a reason to stop looking. It is the reason to look properly. Come up for air whenever you need to. The whale won't.

Forty years she does not need

Start with the rule the whale is breaking. Natural selection is supposed to punish any animal that stops reproducing and carries on eating, because a body that no longer makes copies of itself is, from the genes' point of view, a waste of fish. The theory was set out by Williams in 1957 and Hamilton in 1966, and for most mammals it holds: fertility and the rest of the body wear out together. Menopause, meaning a long life after the ovaries stop rather than a body that simply gives up all at once, has been found in exactly five toothed whales. The killer whale. The short-finned pilot whale. The false killer whale. The beluga. The narwhal. And in us. Not in chimpanzees, and not in elephants, both of which help their relatives a great deal and go on breeding until close to the end.

You can tell, because whales keep the receipts. A female's ovaries hold the scars of past ovulations, so a dead animal's reproductive history can be read from the tissue like a ledger, and in 2018 Samuel Ellis and colleagues read it across 72 cetacean species (Scientific Reports). How it got built had two candidate answers: the menopausal species had either shortened their reproductive lives, "stop early", or lengthened the rest of their lives, "live long". In 2024 the same team compared 18 whale species in Nature, with and without menopause, and the answer was live long. The menopausal species have reproductive spans no shorter than their relatives of the same size. What they have is about forty extra years on the far side, added in a way that does not increase the years in which a mother and her daughter would be breeding at once. Which is, the authors note, the same path the human female took.

So why would evolution pay for forty years of a body that makes no more calves? The answer that has held up best is arithmetic about who

lives with whom. In resident killer whales both sons and daughters stay in their mother's pod for life and mate only with animals from other pods, met in passing. At birth a female is close kin to her mother and almost no kin to the pod's adult males, whose fathers were elsewhere. As she breeds, her daughters stay and have calves that stay, and her sons stay too. Her share of the genes in the water around her rises every year she lives. Rupert Johnstone and Michael Cant put this into demographic models in 2010 and called it kinship dynamics: the older she gets, the more she gains by helping and the less by competing.

But why stop breeding at all, rather than breed and help? Because the older female is closely related to her daughter's calf, her grandchild, while the daughter is much less related to her mother's late calf, a half-sibling, so the daughter has every reason to compete harder for the local food. In 2017 Darren Croft and colleagues showed in Current Biology, from decades of census records, that when the two generations breed together the older female's calf is 1.7 times as likely to die as the younger's. She is outcompeted by her own daughter. That is the mechanism, and it is paid in dead calves.

Now the other side of the ledger, which is what she does with the forty years. Stuart Nattrass and colleagues took more than four decades of census data from the Center for Whale Research and Fisheries and Oceans Canada, 378 grandcalves with known grandmothers, and fitted survival models to them (PNAS, 2019). A calf whose maternal grandmother had died in the previous two years was 4.5 times as likely to die as one whose grandmother was alive, allowing for how many salmon there were. If the grandmother had already stopped breeding, the figure was 6.7. A grandmother who has stopped is worth more to her grandchildren than one who has not, which is the whole argument for the stopping in one number.

Then the sons. Emma Foster and colleagues had already found, in Science in 2012, using 589 individually known whales of which 297 died, that the death of a post-reproductive mother raised the risk of death for her adult sons over thirty in the following year by a multiple that one of my research reports gives as 13.9 and another as 8.3. For adult daughters the figure was 5.4. Either way it is not a tendency. It is a dependency. Michael Weiss and colleagues then priced it, in Current Biology in 2023, on 40 adult females of the Southern Resident population: every surviving son

roughly halves his mother's annual chance of producing another viable calf, and the effect does not fade as he ages. She feeds him, as K16 does, and males with a living post-reproductive mother carry fewer tooth-rake scars from other whales (Grimes and colleagues, Current Biology, 2023). He is also the only way her genes leave the pod, since his calves are born elsewhere and raised at some other mother's expense. I was one of those sons, and I recognise the bargain from the other end.

And she leads. Lauren Brent and colleagues watched the pods travel, hundreds of hours of it, and found in 2015 that post-reproductive females are at the front during collective movement, most of all in years when the Chinook salmon are scarce (Current Biology). What she carries is not muscle. It is a memory of where the fish were the last time the fish were not where they usually are. Mine is the pod line: the order in which a society travels, and who sets it, and whether it changes with the weather. This is what the book means by the Grandmother Clause, the post-reproductive female in front, and this is the animal it was named for.

Three layers, before we leave the water. That her ovaries stop at forty and her body goes on to ninety is measurable, in the tissue. That her sons and grandcalves die more often without her is statistical: hazard ratios, not prophecies, and in a year when the salmon collapse the chance of any female calving halves, matriarch or no matriarch (Eric Ward and colleagues, Journal of Applied Ecology, 2009). That she leads in lean years is culture, in the biologist's sense: knowledge that lives in one animal and is transmitted by being followed. The whales did not confuse the three. Neither should we.

The memory of the herd

Now out of the water, to Amboseli in Kenya, where a loudspeaker is playing an elephant's voice to a family of elephants that cannot see it.

The voice is a contact call, the low rumble a female uses to keep in touch with her relatives, and it belongs either to an animal the family knows or to a stranger. Karen McComb and her colleagues played these calls to 21 families in 2001, with matriarchs aged from 27 to 67 (Science). Families led by old females, the model was run at 55, bunched together and raised their trunks only when the call came from a stranger. Families

led by young matriarchs, modelled at 35, bunched for everyone, burnt the energy and lost the foraging time, and the age of the matriarch predicted how many calves each female in her family produced. A decade later the team did it again with 39 families and the roars of lions, one lion or three, male or female, and the older matriarchs read the threat more accurately and got their families into the right formation faster (Proceedings of the Royal Society B, 2011).

Elephant society is a female gerontocracy, and the old female's asset is a map. When poachers kill the matriarchs for their tusks, which they do first because the tusks are biggest, the map goes with her, and across 218 adult females from 109 groups the families that had lost their elders were still breeding less a decade after the poaching stopped (Gobush and colleagues, 2008, and Goldenberg and colleagues, Current Biology, 2016). She is not decoration. Hold that line. It has a second half.

Because the same species runs a second society, in parallel, and it is nothing like the first. Between ten and fifteen a young male leaves his mother's herd, or is pushed out of it by the older females, and enters the world of bulls. That world runs on musth, a state in which a male's testosterone rises from a resting level of around 2 nanograms per millilitre to 20 or 40 (Poole, 1987 and 1989), his temporal glands stream, he loses weight and becomes extremely dangerous. How long he can hold it depends on his age. Poole's medians: two days a year for males of 16 to 25, thirteen days at 26 to 35, 81 days at 46 to 50, falling back to 54 days at 51 to 60. So the arithmetic of who fathers the calves is set by who can stay in musth longest, and in Amboseli, over 22 years and 119 calves matched against 89 sampled adult males, paternity peaked between 45 and 53 (Hollister-Smith and colleagues, Animal Behaviour, 2007). The other column: males who were not in musth still sired about 20 per cent of the calves, by sneaking. Whether musth is a state of suffering is contested: 60 wild bulls in Kaziranga, India, showed no rise in stress hormones during musth (Ghosal and colleagues, PLoS ONE, 2013), while four bulls in captivity did (Ganswindt and colleagues). Musth is measurable. Paternity is statistical. What follows is culture, in the sense that matters.

In the late 1990s the Kruger National Park in South Africa culled elephants and sent the orphaned young males to Pilanesberg National Park, without any older bulls. The young males reached adolescence with no adults of their own sex, entered musth years early, stayed in it far too long,

and killed more than 100 white rhinoceroses. The park brought in six older bulls. Nothing else changed. The young males' musth shut down, the killing stopped, and Slotow and colleagues reported it in Nature in 2000 under the title "Older bull elephants control young males". What the old bulls carried was not a weapon but a signal that said: you are not the biggest thing here.

So here is the "natural" elephant. A female society run on the memory of the oldest mother, and a male society in which the presence of old males switches off the hormones of young ones. Anyone who wants the elephant to prove that nature is matriarchal has to explain the bulls. Anyone who wants the bulls to prove that males run on testosterone has to explain Pilanesberg, where the testosterone ran on the absence of older males.

Sisters who are not sisters

Two apes are equally our nearest relatives. They split from each other only one to two million years ago and are separated today by the Congo River, and they disagree about the one thing this book is about.

The chimpanzee is the ape most people mean when they say "our cousin". Males weigh about 42.7 kilograms to females' 33.7, a ratio of 1.27 (Smith and Jungers, 1997). Every adult male outranks every adult female, and males stay in the group they were born in and form alliances that run it. Those alliances also raid. Richard Wrangham's imbalance-of-power hypothesis (1999) describes what happens when a large party of males meets a lone neighbour: at a typical advantage of eight to one, they kill him at almost no risk to themselves. Wilson and colleagues pooled 18 chimpanzee communities and four bonobo communities in Nature in 2014, 152 killings observed or inferred, and found that the killing did not track human disturbance, which was the critics' explanation. The anthropologist Brian Ferguson and the writer John Horgan still call this an artefact of what humans have done to the forests. The primatologists mostly do not. The gorilla, for completeness, is the harem, at a mass ratio of 2.24, with infanticide by incoming males accounting for up to 21 per cent of infant deaths (Robbins and colleagues, PLoS One, 2013).

The bonobo, at a ratio of 1.36, is if anything slightly more dimorphic

in body mass. And in the bonobo, females outrank males. For decades the argument was whether this was true dominance or something politer, and in 2025 Martin Surbeck and colleagues settled most of it with 30 years of records from six wild communities and 1,786 conflicts between a female and a male (Communications Biology). Females won 61 per cent of them, 1,099 fights, and on average outranked 70 per cent of the males. They did not do it one to one. In 85 per cent of the coalitions the researchers observed, females ganged up on a male, loudly and together, and the females who did it most often ranked highest. The heavier animal loses because the lighter ones arrive in threes.

What makes this remarkable is that bonobo females are not sisters. Like chimpanzee females, they leave the group they were born in at maturity, so the adult females of any community are unrelated immigrants. They build the alliance from nothing, and the material is sex: females rub genitals with one another, most often in tense feeding situations or after a fight, and the act raises oxytocin in their urine more than mating with a male does (Moscovice and colleagues, Hormones and Behavior, 2019). A male's rank, meanwhile, depends on his mother. Bonobo mothers escort their sons into adult society and their sons' status rises and falls with theirs. The distributions overlap: males won 39 per cent of those fights, and in some communities, Ekalakala and Bompusa West are named, a male held the top rank for a year or more, usually with a high-ranking mother behind him. Surbeck's own phrase for female dominance is that it is "by no means the rule" in every encounter. It is the rule in the average. That is a statistical claim, and the mechanism is coalition, and the coalition is culture in the sense that it has to be rebuilt by every generation of immigrants.

Which of these is the ancestor? The orthodox answer for thirty years was the chimpanzee. Then the fossils spoke: *Ardipithecus ramidus*, 4.4 million years old, has canine teeth barely larger in males than in females, a ratio of 1.06 to 1.13, smaller than the bonobo's (Suwa, Lovejoy and colleagues, PNAS, 2021). Neither the warlord nor the sisterhood. Something with less to prove.

I could stop at the apes, but I promised you every direction at once. Female spotted hyenas outrank males and give birth through an erectile pseudo-penis, which kills about 60 per cent of first litters (Holekamp and Strauss, PNAS, 2019 and 2022). Among the lemurs of Madagascar the

adult sex ratio, across 253 populations of 121 primate species, is the best predictor of who gives way to whom (Kappeler, Biological Reviews, 2022). Several routes to a female at the top, none of them the same, and one of them frankly horrible. Nature is not making a recommendation.

What the ratio decides

Then what does decide it? For seventy years the answer was a fruit fly.

In 1948 Angus Bateman put marked fruit flies in jars and counted their offspring, and reported that a male's reproductive success rose with the number of his mates and a female's did not, and that males varied more in success than females. He explained both by the size of the gametes: sperm is cheap and eggs are dear, so the egg-maker is the scarce resource and the sperm-maker competes for it. Trivers extended the logic to parental care in 1972. Coy females, competitive males, weaponry and dimorphism as the visible result. It was clean, it was memorable, and it was the entire theory of sex roles taught to my generation.

In 2012 Gowaty, Kim and Anderson repeated Bateman's experiment as exactly as they could and found no evidence of sexual selection at all (PNAS). The markers Bateman had used to tell offspring apart were lethal mutations that killed some of them before they could be counted, and the deaths fell unevenly. The founding experiment did not replicate. What did survive was the average: Janicke and colleagues pooled up to 102 effect sizes from 64 to 66 species in Science Advances in 2016 and found that sexual selection is, on average, stronger on males, with enormous variance. In pipefish, jacanas and some bush crickets the gradient is steeper for females. Selection tracks whichever sex is limiting at the time. It is not bolted to the sperm.

What decides which sex is limiting? Kokko and Jennions gave the answer in 2008 (Journal of Evolutionary Biology), and it begins with a fact so simple it was overlooked: every offspring has exactly one mother and one father, so if there are fewer adults of one sex, that sex does better per head, by arithmetic. What matters is the adult sex ratio, and what sets it is which sex dies more, and that depends on ecology, not gametes. Trivers's idea that the sex which has already invested more should go on investing they called the Concorde fallacy, after the aircraft nobody could stop

funding. If females die young and males are the surplus sex, a male's best move is not to compete for a scarce mate but to stay and raise the young he has.

The birds proved it. Eberhart-Phillips and colleagues followed 6,119 individual shorebirds in six wild populations (Nature Communications, 2018). The sex ratio at hatching was near even everywhere. What differed was who survived the first year. In the snowy plover, young males survived 51 per cent better than young females, the adult population ended up 63 per cent male, and the whole arrangement of care flipped: males did all the parenting, and the scarce females laid a clutch, left it to the male, and went off to find another. Where the ratio was even, the pair shared. Same genus, same eggs, same gametes. The ratio decided.

One more rule and then the humans. Social monogamy, a male and a female living as a pair, is rare in mammals, about 9 per cent of species against 90 per cent of birds, and in every one of the 61 times it evolved, Lukas and Clutton-Brock found across 2,545 species in Science in 2013, the ancestral females were solitary, intolerant of each other and spread too thinly for a male to guard more than one. Paternal care came afterwards, in 127 of 226 monogamous mammals, as a consequence of pair-living rather than its cause.

Now us, on the same axes. Men are on average 15 to 20 per cent heavier than women and 7 to 8 per cent taller, which puts us in the middle of the primates, nowhere near the gorilla. But the composition is another matter. In 468 people aged 18 to 88, scanned whole-body by MRI, men carried 33.0 kilograms of skeletal muscle and women 21.0, 38.4 per cent of body mass against 30.6, and the gap was 40 per cent in the upper body and 33 in the lower (Janssen and colleagues, Journal of Applied Physiology, 2000). Overlap exists at the tails and is small at the shoulders. On the animals' rules that is the body of a species in which males fought males, and it is also the species with the slowest-growing brains in the class, which cannot be fed by one mother. Kaplan and colleagues (2000) argue that a skill-intensive diet forced a long helpless childhood and with it a father who provisions and a grandmother who helps (Hawkes, 2003). Moderate polygyny, heavy paternal care, three generations feeding one child. No other mammal does that combination, and it is the reason chapter 2 exists.

Who moves the pod

So far the animals have told us who ends up at the front. The last question is how the group decides to follow, and here the picture is more democratic than either shelf would like.

In Kenya, Strandburg-Peshkin and colleagues put GPS collars on 25 wild olive baboons, 80 per cent of the adults and subadults in the troop, recording every animal's position once a second (Science, 2015). The old assumption was that the dominant males set the route. They do not. A baboon is more likely to follow a movement when several animals start it together, regardless of their rank, and when two subgroups pull in different directions the troop splits the difference if the angle between them is under about 90 degrees and picks one, usually the one with more initiators, if it is larger. Rank is not the currency. Agreement is.

African wild dogs in Botswana vote by sneezing. Before a hunt the pack rallies, and Walker and colleagues counted, across 34 rallies (Proceedings of the Royal Society B, 2017), that when a dominant animal started the rally about three sneezes were enough to launch the pack, and when a subordinate started it, about ten. A weighted vote, then, in which rank buys a discount but enough subordinates can outvote anyone.

One more finding and then the tool. Smith and colleagues compared leadership in eight mammal societies, elephants, dolphins, chimpanzees, lions, meerkats, zebras, hyenas and capuchins, with eight small-scale human societies from the Ache to the Tsimane (Trends in Ecology and Evolution, 2016). In both, leadership of movement is shared and mild, and leadership in conflict with other groups is concentrated in a few. The differences were three. Animal leaders lead in everything at once, human leaders lead in one domain and follow in another. Animal leaders coerce, human leaders in small societies mostly persuade. And humans share food by rule, which almost no other mammal does. We took the mammalian pattern and specialised it, softened it and fed it.

Here, then, is the tool the animals hand over, and it is not a verdict. It is a short list of the routes by which a society with two sexes in it ends up with somebody at the front. The ratio: who is scarce, and therefore has leverage, as with the plovers and the lemurs. The coalition: who arrives in threes, as the bonobo females do. The memory: who has seen the drought before, as the matriarch and the old whale have. The older animal of the

same sex in the room: what Pilanesberg lacked. And the quorum: how many sneezes it takes, and whether rank buys a discount. Ask of any group, a couple, a company, a parliament, which of those routes put the person at the front there, and whether the order changes when the fish run out. That question is the pod line, and I intend to ask it at every scale in this book. My mother's generation asked whether a woman could be at the front. The whales, the elephants and the bonobos answer with a shrug. Of course. The question they ask back is harder: by which route, at what cost, and to whom.

The foragers asked it first, and they wrote nothing down.

THE STRONGEST CASE AGAINST

The strongest case against this chapter is that the whale is a whale.

Begin with the animal that should have menopause and does not. The long-finned pilot whale has the same social structure as its short-finned sister: both sexes stay in the natal pod, both mate outside it, so the kinship dynamics that supposedly explain the killer whale's menopause apply to it in full. Its females breed until they die. McCormack, Arbuckle and Nichols traced parentage in 25 pods off Cape Breton (Behavioral Ecology, 2023) and found no sign of the mother-daughter conflict that drives the killer whale story: older females help their daughters breed and pay no calf-mortality penalty for breeding alongside them. So the grandmother's help alone was never enough to build the Grandmother Clause. It took a specific and rather nasty cost, and a species can have the whole matrilineal apparatus and skip the clause entirely. The older critics, Cohen in 2004 and Levitis, go further: the post-reproductive measure depends on how you count, and for the boundary cases, chimpanzees and elephants, the argument is not settled.

Then the size of the effect. Ward and colleagues (2009) found that in resident killer whales the abundance of Chinook salmon swamps everything: in a bad year the chance of a female calving falls by half whether or not her mother is alive to lead. Møller and Jennions went through 43 published meta-analyses in ecology and evolution in 2002 (Oecologia) and found that the biological factor under study explained between 2.51 and 5.42 per cent of the variation, which is the honest ceiling on what any of the biology in this chapter explains. A matriarch's age

predicts her family's calves on average, not on a given afternoon. And the chapter's flagship figure cannot yet be stated: the same 2012 paper gives, in two research reports on my desk, an 8.3-fold and a 13.9-fold rise in a son's mortality. The salmon scene that opens the chapter is sourced, at present, to a post on social media.

Then the transfer problem, which is the real one. Smith and colleagues' 2016 comparison is the best evidence that human leadership is not the mammalian kind: it is specialised by domain, it runs on persuasion rather than coercion, and it is embedded in food-sharing rules that no other mammal has. The animals are least like us exactly where this book wants to use them. A bonobo coalition holds because the females rub genitals in tense moments. A matriarch's authority holds because a family that gets lion roars wrong loses calves. Neither mechanism is available to a board of directors, and the one thing the chapter has shown for certain, that nature does everything in every direction, is as much an argument for ignoring nature as for consulting it. If the chimpanzee and the bonobo, at the same distance from us, give opposite answers, then the choice of which cousin to cite is made for other reasons, and calling the result "natural" adds nothing but borrowed authority. That is the naturalistic fallacy in a wetsuit.

My reply, briefly, is that I have not asked the animals for permission. I have asked them for the list of routes, and the list is the same whether or not you like any item on it. But the objection stands as a limit on everything that follows: what the whale shows is that a society can put an old female at the front. What it cannot show is that ours should.

IN PRACTICE

This chapter is about how to read the word "natural" when it is used on you, at a dinner table, in a meeting, in a headline. So the practice is a way of reading, not a way of living.

If someone tells you that an arrangement between men and women is natural, ask which species. If they mean the chimpanzee, the bonobo stands at the same distance and says the opposite. If they mean the gorilla, they are describing a harem with a 21 per cent infanticide rate. If they mean the whale, they are describing a grandmother who feeds her son for life. The failure mode is that this becomes a party trick that stops the conversation, and the counter-evidence is Janicke's average: across the animal

kingdom, sexual selection really is stronger on males, so "which species" is a question, not a refutation.

If you are the oldest person in a group, of either sex, the elephant and whale data suggest that your value rises in the lean years and is nearly invisible in the fat ones, so a group which drops its elders in a good year tends to find out what they knew in a bad one. The failure mode is the matriarch who reads every unfamiliar rumble as a threat because she is not, in fact, old enough yet, or who leads the pod to where the salmon were in 1990. Ward's salmon data are the counter-evidence: in a real collapse, nobody's memory is enough.

If your group has no older members of one sex, watch the younger members of that sex. Pilanesberg is one park, one species and one decade, a warning rather than a finding. But the young bulls' hormones did not regulate themselves, and the fix was presence, not punishment. The failure mode is to read this as "young men need a strong man over them", which is not what the study says: the older bulls did not fight the young ones. They stood there. The counter-evidence is that wild Kaziranga bulls in musth showed no stress at all, so musth is not, in itself, a pathology.

If you are the smaller party in a contest, physically or otherwise, the bonobo suggests the route is not to become larger but to arrive in threes. This holds in 30 years of bonobo records, and it holds better the more often the coalition is rehearsed. The failure mode is the coalition that exists only when needed, which is to say never: bonobo females build theirs daily, in peacetime. The counter-evidence is that in two of the six communities a male held the top rank for years, with his mother behind him, so the coalition is a tendency and not a guarantee.

If the ratio in your group is skewed, expect the scarce sex to have leverage whether or not anyone intends it. In the plovers the surplus sex did all the care and the scarce sex left. In the lemurs the ratio, not the muscle, predicts who gives way. The failure mode is to treat leverage as a personality: the plover male is not more nurturing, he is more numerous. The counter-evidence is human culture's habit of turning the dials: a skewed ratio in a room tends to produce leverage only where no rule, norm or pay scale is already producing something stronger.

If you are deciding something as a group, the baboons and the wild dogs suggest that agreement among several starters moves a group better than rank does, and that a quorum with a rank discount is what most

social mammals actually run. That tends to work when the group can see who is starting and count them. The failure mode is the herd waiting for a lead mare who does not exist, and the counter-evidence is Smith's finding that in conflict with outsiders every society in the sample, animal and human, concentrates leadership in a few. Democracy is for the walk to the river. It is not, apparently, for the war.

And if you are tempted to conclude anything about your own household from any of this, hold the three layers apart first. The whale's menopause is measurable. Her sons' survival is statistical. Her leadership is culture. What the human equivalents are, in a flat in a city with a tax code, is the work of the chapters that follow. What the animals can do is stop you saying "natural" as though it settled anything.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. Menopause in five toothed whales and humans, built by adding about forty years rather than subtracting fertile ones (Ellis 2024). Kinship dynamics as the reason (Johnstone and Cant 2010) and reproductive conflict as the cost (Croft 2017). The grandmother effect on grandcalves (Nattrass 2019) and the mother effect on adult sons (Foster 2012). Matriarch age and family reproductive success in elephants (McComb 2001). Older bulls suppressing musth in young males (Slotow 2000). Bonobo female coalitions overriding a size disadvantage (Surbeck 2025). Adult sex ratio, not gamete size, as the driver of sex roles (Kokko and Jennions 2008, Eberhart-Phillips 2018). Shared decision-making in collective movement (Strandburg-Peshkin 2015).

Decoration. The K16 salmon scene. The sneezing wild dogs. The hyena and the lemur. The Ardipithecus canines.

Contested. Whether menopause is adaptive or an epiphenomenon (Cohen 2004, Levitis, versus Ellis 2024). The size of Foster's son effect (8.3 or 13.9). Whether chimpanzee killing is adaptive or a human artefact (Wilson 2014 versus Ferguson and Horgan). Whether the ancestor was chimp-like, bonobo-like or neither (Suwa 2021). Whether bonobo females are "dominant" or "co-dominant" (Parish 1996 versus Surbeck and Furuichi). Whether musth is a stress state (Ghosal versus Ganswindt).

NUMBERS TO CARRY

- Menopausal whales gained about 40 years of life, not shorter reproductive spans: 18 species compared (Ellis et al. 2024, Nature, via AN-01).
- A grandcalf whose grandmother died in the previous two years is 4.5 times as likely to die, 6.7 times if she was post-reproductive: 378 grandcalves (Nattrass et al. 2019, PNAS, via AN-01).

- An adult son over 30 is 8.3 or 13.9 times as likely to die in the year after his post-reproductive mother's death, a daughter 5.4 times: 589 whales, 297 deaths (Foster et al. 2012, Science, via AN-01 and AN-06, which disagree).
- Bonobo females won 61 per cent of 1,786 conflicts with males and 85 per cent of female coalitions targeted males: 30 years, six communities (Surbeck et al. 2025, via AN-03 and AN-06).
- Biological factors explain 2.51 to 5.42 per cent of the variance across 43 ecological meta-analyses (Møller and Jennions 2002, Oecologia, via AN-09).

THE LIBRARY

- AN-01 Whales: matrilineal societies and menopause. The chapter's spine: Ellis, Johnstone and Cant, Croft, Nattrass, Foster, Weiss, Grimes, Brent, Ward, McCormack.
- AN-02 Elephants: matriarchs and bull society. McComb, Goldenberg, Poole, Hollister-Smith, Ghosal, Slotow. Wittemyer 2007, Wasser 2008 and Allen 2020 cut.
- AN-03 Great apes: four models of sex relations. Dimorphism figures, Surbeck, Wilson, Suwa, Moscovice.
- AN-04 Pair-bonding mammals. Lukas and Clutton-Brock. Opie 2013, Murdock 1967, Henrich 2012 and the voles are held for chapters 2 and 6.
- AN-05 Cross-species rules: what predicts sex roles. Bateman, Gowaty, Janicke, Kokko and Jennions, Eberhart-Phillips, Janssen, Kaplan.
- AN-06 Female-dominant and female-led species. Hyena, lemur, the second Foster figure, the second McComb sample. Meerkat and mole rat cut.
- AN-09 How mixed-sex mammal groups decide and lead. Strandburg-Peshkin, Walker, Smith, Møller and Jennions. Silk 2003 and the horse studies held for chapter 7.

CHAPTER 2

The Campfire Contract

A woman of about sixty is kneeling in dry ground north of Lake Eyasi, in northern Tanzania, working a digging stick down beside a vine. The stick is a length of hardwood with the tip hardened in a fire, and she has been using it since mid-morning. The tuber she is after sits half a metre down. Her arms are thin and the muscle on them is laid on like cord, because this is what her arms have done most days for fifty years. A child, her daughter's youngest, sleeps on a cloth in the shade of a thorn bush, and when it stirs she talks to it without looking up.

By late afternoon she has a bundle that weighs more than the child. She carries it back to camp on her head. The men have been out since dawn with bows, and one of them has a dik-dik across his shoulders, which is a good day, and the others are carrying nothing, which is the ordinary day. The tubers go into the coals. A share of them goes to the children of the woman who dug beside her, and a share goes to her own grandchildren, and a smaller share goes to her, and nobody counts.

Sit a little way back from the fire, where the smoke is thinner, and watch who hands what to whom.

The dik-dik is cut and passed around by rules that everyone knows and nobody states. The hunter is not praised. If anything he is teased, and he lowers his head and agrees that it was a small animal and a lucky shot. Somebody's brother-in-law has not brought anything in for a week and is still eating. A young man who tried to keep back a piece last month was talked about for days and has not tried it since. The old woman who dug the tubers is not thanked either. Her daughter's children ate, and that was

the point of the digging, and the point was understood.

As the light goes, the talk changes. In daylight it was about who owes whom, who did what, who is lazy. Now it is stories, the old ones, about people who are dead and animals that speak, and the grievances are set down like tools at the end of a working day.

There is a ledger here. Nobody wrote it. It records who works at what, who eats first, who decides, who moves at marriage and who stays, who holds the baby and who holds the bow, and it has been kept, by nobody, for something like two hundred thousand years. Every camp keeps its own copy and every copy is different, and yet, when you lay the copies side by side, the same entries keep appearing in the same hand. This chapter is about those entries, and about what happened to the ledger when people stopped moving and started to own things.

Who ate what

I should say at once what kind of evidence this is, because there is no film of the Pleistocene. What there is comes in three kinds, and the three keep different books. Bones, teeth and ancient genomes are the physiological layer, measurable and mute. The living foragers studied in the twentieth century, the Hadza of Tanzania, the Ju/'hoansi of the Kalahari, the Aka and Mbendjele and Efe of the Congo forests, the Agta of the Philippines, the Aché of Paraguay, are the behavioural layer, statistical and overlapping, and they are not Stone Age people but modern people with a particular economy. And the meanings the foragers themselves give to what they do, the myths and initiations and taboos, are the cultural layer, and they vary from valley to valley. When I write that foragers did something, I mean that the entry appears in most copies of the ledger, not that every camp kept it.

Chapter 1 laid out the bodies, so I take only one figure from that layer here. The male surplus of muscle is not spread evenly but concentrated in the arms, where William Lassek and Steven Gaulin's 2009 review in Evolution and Human Behavior put it at 75 to 90 per cent more than women's (DH-01). Physiology hands each sex a different tool. What the ledger records is how the tools were used, and here the cultural layer sits on the physiological one with a consistency that surprised the people who first

counted it. In George Murdock and Caterina Provost's 1973 tally of 50 activities across 185 societies, cooking was women's work in 183 of the 185, while smelting, metalworking and the hunting of large game were men's work in 90 to 100 per cent of them (DH-07). Not one of the 185 assigned large-game hunting to women alone (DH-07). Judith Brown's explanation, offered in 1970 in American Anthropologist, still carries most of the weight: the tasks that went to women were the ones that could be interrupted, done near camp and done with a child on the hip, and the tasks that went to men were the ones that could not (DH-07).

That is the average. The overlap is large and it is the interesting part. Among the Martu of Australia's Western Desert, women hunting sand monitors bring in the majority of the calories eaten in camp (DH-02). A 2023 survey in PLOS ONE by Abigail Anderson and colleagues found women hunting in 50 of 63 foraging societies with usable data, and hunting large game in a third of those, and the paper was taken up everywhere as the end of man the hunter (DH-02). The rebuttal came in 2024 from Vivek Venkataraman and colleagues, who could not locate about 35 per cent of Anderson's societies in the D-PLACE database and argued that occasional hunting had been counted as the household's plan (DH-02). Both things are true. Women hunt, sometimes a great deal. The division of labour by sex is also real, and it is a division of averages, not of persons. At Wilamaya Patjxa in the Peruvian Andes, a woman of 17 to 19 was buried about 9,000 years ago with five projectile points and a butchering kit, and Randall Haas's 2020 review in Science Advances of 27 comparable burials found 11 female and 16 male, which he read as female participation in big-game hunting of 30 to 50 per cent (DH-04). The bones are the physiological layer. The reading of them is the cultural layer of our own decade, and it is still being argued.

What the ledger does not record is a leisured sex. Herman Pontzer's team measured 30 Hadza adults, 17 women and 13 men, with doubly labelled water and found the men burning 2,649 kilocalories a day to the women's 1,877 (DH-02), a difference in body size and walking distance rather than in effort. A 2024 study of time use by Deb and colleagues in Evolution and Human Behavior found leisure divided within forager households with what its authors called remarkable equality (DH-02). In the wealthy countries where I live and write, women do 70 to 90 per cent of household chores (DH-02). The foragers did not arrive at that number,

and they did not avoid it by writing a rule against it.

Nor was the camp built on men's kinship. Mark Dyble's 2015 paper in Science modelled 1,863 Mbendjele and 4,055 Agta pairs of co-residents and found that a camp in which both sexes have a say in where the household lives ends up with about 12 per cent of its adults unrelated to anyone else, against 0.6 per cent when one sex decides, and the observed Agta figure was 11.2 per cent (DH-03). The arithmetic of a mixed camp is the arithmetic of two people deciding. The Paranan farmers next door, who had settled, came in at 4.2 per cent (DH-03).

Insulting the meat

One Christmas in the late 1960s, Richard Lee, the Canadian anthropologist who had spent years with the Dobe Ju/'hoansi, then written as the !Kung, in the Kalahari, bought the fattest ox he could find to thank the camp that had put up with him. He expected pleasure. He got, for days, a chorus of complaint: the animal was a bag of bones, it would feed nobody, he had been cheated. When the ox was butchered and turned out to be layered with fat, the complaining went on anyway. Lee wrote it up in 1969 in Natural History, and an old man explained the rule to him. When a young man kills much meat he comes to think of himself as a chief, and this the camp refuses, because someday his pride makes him kill somebody. So his meat is called worthless, and in this way, the old man said, "we cool his heart and make him gentle" (DH-03).

Christopher Boehm, the anthropologist at the University of Southern California, made the cooling of hearts into a theory. In 1993 in Current Anthropology he compared 53 forager and tribal communities and found the same thing in all of them: not the absence of a dominance hierarchy but its reversal, a coalition of the rank and file holding down whoever tried to climb (DH-03). The sanctions run in a fixed order, and the order is a ledger entry in itself. Gossip first. Then ridicule. Then ostracism. Then, rarely, execution (DH-03). Boehm called it reverse dominance hierarchy, and the phrase carries a claim about physiology as well as culture. The weapons that made a strong man dangerous, the spear and the poisoned arrow, made him equally killable in his sleep, and a coalition of the weaker could always outnumber him. Richard Wrangham's 2019 argument is

that this went on long enough to change the species: the men most given to impulsive aggression were executed, generation after generation, until the skulls softened. Robert Cieri's 2014 study of more than 1,400 crania in Current Anthropology found brow ridges and faces becoming lighter across the last 200,000 years (DH-03), which is the physiological layer of a cultural habit, if the argument holds. Note the conditional. Wrangham's is a theory, and the crania could be recording other things.

What the levelling looked like from inside was recorded by Polly Wiessner, who in 2014 in PNAS reported 174 conversations she had noted among the Ju/'hoansi over decades, 122 by day and 52 at night (DH-03). By day, 31 per cent of talk was economic, hunting plans and gathering, and 34 per cent was complaint, criticism and conflict, the grinding of the mill (DH-03). At night, by the fire, 81 per cent was stories, myth and healing, and conflict fell to almost nothing (DH-03). The levelling was done in daylight, by talk. The repair was done in the dark, by other talk. And behind both stood the exit. A household that could not bear a camp could walk to another one, because the Ju/'hoansi network of hxaro gift partners gave every adult standing in several camps at once (DH-03). Ridicule works when the ridiculed can leave and when the ridiculers can be left.

James Woodburn's 1982 distinction in Man between immediate-return economies, which eat today what they took today, and delayed-return economies, which store and therefore own, is the hinge on which the whole ledger turns (DH-03). Where there is nothing to inherit there is nothing to guard, and where there is nothing to guard a strong man is only a strong man. Monique Borgerhoff Mulder's 2009 comparison of 21 societies in Science put the Gini coefficient of wealth among foragers at about 0.17, against 0.35 and above among farmers and herders (DH-03), and found that a child born into the top tenth of an agricultural society was 11 times likelier to stay there than a child born at the bottom (DH-05). The foraging bands did not achieve a rough equality between the sexes by believing in it. They achieved a rough equality among adults by preventing anyone from accumulating, and women were carried along inside that.

Not always, and not all the way. The levelling I have described was among men, mostly, over meat and pride. What levelled the relation between the sexes was something else, and it had other hands.

Other hands

Sarah Blaffer Hrdy, the primatologist at the University of California, Davis, put the cost of a human child at roughly 13 million calories from conception to independence (DH-03), and made the point that no ape mother could pay it alone. Humans belong to the roughly 3 per cent of mammal species that breed cooperatively (DH-01), and Hrdy's argument in Mothers and Others in 2009 is that a mother's willingness to hand her infant to other hands, and the willingness of other hands to take it, is the oldest entry in the ledger and the one on which the rest was written (DH-08). The entry does not read "women look after children". It reads "children are looked after by several people, most of them female, and the mother does about half".

The numbers are startling to a parent in a flat. Paula Ivey's 2000 study of the Efe of the Ituri forest counted an infant passing through the hands of 14.2 caregivers a day on average, with a range of 5 to 24, and by 18 weeks the child spent 60 per cent of daylight with someone other than its mother (DH-08). Among the Aka, whom Barry Hewlett followed in 15 families, fathers were within arm's reach of their infant for 47 per cent of the day and held it for 57 minutes a day against the mother's 490 (DH-08), a real share and still an eighth of hers. Karen Kramer's work with Maya and Pumé families found that when other carers took 10 per cent of the load, the mother's own care fell by 25 per cent and she saved about 165 kilocalories a day (DH-08). Courtney Meehan measured grandmothers saving a mother up to 216 kilocalories over nine hours (DH-08). The birth interval among the Ju/'hoansi that Melvin Konner and Carol Worthman described in Science in 1980 was about 44 months, and a woman bore about 4.7 children (DH-08), and the interval was held there by nursing on demand, day and night, for close to three years.

Which brings me to Kristen Hawkes. Hawkes, the anthropologist at the University of Utah, had watched Hadza grandmothers dig tubers through the 1980s and 1990s and noticed that the old women's returns went to their daughters' weaned children, so that a daughter with such a mother could wean sooner and bear again sooner (DH-01, DH-02). With James O'Connell, Nicholas Blurton Jones, Helen Alvarez and Eric Charnov she published the grandmother hypothesis in PNAS in 1998: that the human female's long life after her last birth, which no other prim-

ate shares, was selected because a woman who fed her grandchildren left more descendants than a woman who kept bearing (DH-01). The *Grandmother Clause* of the last chapter is this, the post-reproductive decades as a working part of the household rather than a remainder. It is a hypothesis, and the competing account, Owen Lovejoy's 1981 claim in Science that it was fathers provisioning mates who paid the bill, is still on the shelf (DH-01). They are not exclusive. A human infant is expensive enough for both to have been true.

Rebecca Sear, the anthropological demographer in London, tested the grandmother directly. With Ruth Mace in 2008 she reviewed 45 populations, historical and contemporary, in which the presence of a relative could be matched to a child's chance of surviving (DH-08). The maternal grandmother improved survival in nearly every population where the effect was measured. The father improved it in only about a third. The paternal grandmother's effect was variable and in some populations negative (DH-08). That is a statistical finding across populations and says nothing about any one father, and it is the most durable number in this chapter. The behavioural layer is that the extra hands mattered most when they belonged to the mother's mother. The cultural layer is where those hands lived, and that turns out to depend on who moved at marriage, which is the next entry but one.

Two halves of a village

Because the ledger was never written, a society could keep it either way round. Most kept it with men at the centre, and some did not. The Minangkabau of West Sumatra, more than 4 million people, trace descent and pass ancestral land through women (DH-06). The Khasi of Meghalaya give the house and its property to the youngest daughter, the khatduh, and for most of their history barred women from the village council (DH-06). The Mosuo of Yunnan practise tisese, the walking marriage, in which a man visits at night and returns to his mother's house by morning, and brothers help raise their sisters' children (DH-06). Among the Haudenosaunee, clan mothers named the sachems and could remove them, and held a veto on war (DH-06). Nowhere in this list is there a matriarchy, a mirror image of the rule of men. The search for one, from Johann Jakob

Bachofen in 1861 to the Goddess of Marija Gimbutas, has been taken apart by Cynthia Eller and by the excavators of Çatalhöyük, who reported zero evidence of one (DH-05, DH-06). Joan Bamberger's 1974 reading has lasted: the myths of a lost rule of women, told across the Amazon and Melanesia, are charters for the rule of men, stories about why women had to be put down (DH-06).

What the matrilineal societies do show is what changes in the body when the ledger is turned. Among the Mosuo, Adam Reynolds and Siobhán Mattison's team measured inflammation in 2020 in PNAS and found that in the patrilineal villages 8.4 per cent of women had chronic inflammation against 3.2 per cent of men, while in the matrilineal villages the figures were 3.6 per cent of women and 6.5 per cent of men (DH-06). Physiology follows the entry. And Uri Gneezy, Kenneth Leonard and John List's 2009 experiment in Econometrica found 50 per cent of Maasai men and 26 per cent of Maasai women choosing to compete for a prize, and among the Khasi 54 per cent of women and 39 per cent of men (DH-06). The willingness to compete reversed with the descent rule. Not entirely, and not in every replication, since Chakravarty's Dimasa showed no gap and the Mosuo and Yi comparison found the gap in risk-taking smaller but still there (DH-06). The Overlapping Bells move when the rule moves. They do not swap places.

Complementarity is the word anthropologists reach for, and it is not the same word as equality. Martin Whyte's 1978 study of 93 societies found no such thing as a single status of women, but ten independent dimensions that did not travel together, so that a society could give women property and ritual authority and no voice at all in public matters (DH-07). In his sample, 88 per cent of societies barred women from formal political leadership (DH-10). What they had instead was a parallel institution. The Sande society of Sierra Leone and Liberia initiates girls, teaches them, judges them, and defends their interests against the men's Poro, which does the same the other way (DH-10). Among the Gê of central Brazil the village is two halves, the men's plaza at the centre and the women's houses around it, and each half buries the other's dead (DH-10). Even the menstrual hut, the standard example of women's subjection, turns out to keep a ledger. Beverly Strassmann counted 736 consecutive days of hut visits among the Dogon of Mali and ran hormone assays on 93 women, and the women went to the hut for 86 per cent of their true

menses and never while pregnant (DH-10). Her reading is that the hut was an honesty device for husbands. Chris Knight's reading is that it was a coalition of women. The hormones cannot tell you which, and neither can I.

Who moved and who stayed

Then people stopped moving. The ledger did not stop, but it began to be kept in a different place, and for the first time parts of it were written.

The genomes give the residence rule before anything else does. Across the world, the maternal lineages in living people go back roughly twice as far as the paternal ones, in a 2014 study of 623 men from 51 populations by Sebastian Lippold and colleagues (DH-12), which means that across deep time fewer men than women left descendants. And in 2022 Laurits Skov's team in Nature read the genomes of 11 Neanderthals from Chagyrskaya Cave and two from Okladnikov Cave in the Altai, about 54,000 years old, a father and his teenage daughter among them, and found more diversity in the maternal lines than the paternal, the signature of women moving between groups of 10 to 20 people (DH-12). Patrilocality, the wife going to the husband's people, is older than our species.

The Neolithic did not invent it. It fixed it. At Gurgy in France, Maïté Rivollat's 2023 paper in Nature sequenced 94 individuals buried between 4850 and 4500 BCE and drew a pedigree of 63 people across seven generations, every link running through a father, the adult daughters missing because they had married out, and the wives unrelated to each other and non-local by their strontium (DH-05, DH-12). The reports disagree on how many were sampled, 110 or 128, and I carry both. In the Lech valley of southern Germany, Alissa Mittnik's 2019 study in Science found nine of ten parent and child pairs on the farms were father and son, and the mothers had come from up to 350 kilometres away, buried with wealth beside unrelated people buried with nothing (DH-12). At Hazleton North in Britain, a single man 5,700 years ago had children by four women and the tomb was organised around six of his sons (DH-12). Read that against the forager camp with 12 per cent of its adults unrelated and the shift is visible in the bones: from a camp two people chose to a farm one lineage kept.

And then the Y chromosome collapses. Monika Karmin's 2015 study in Genome Research of 456 Y chromosomes from 110 populations found that in the millennia around 5000 to 3000 BCE the number of men leaving paternal lines fell so far that in some regions roughly 17 women reproduced for every man whose line survived (DH-05, DH-12). The average man alive in 4000 BCE has no living descendant in the male line. Two explanations are on the table and the reports I am working from do not settle them. Tian Chen Zeng, Alan Aw and Marcus Feldman argued in 2018 that patrilineal clans fought and the losers' men died (DH-12). Léa Guyon, Nina Marchi, Evelyne Heyer and Raphaëlle Chaix answered in 2024 that no massacre is needed: rich clans split and poor clans fail to marry, generation after generation, and the arithmetic reaches 17 to 1 in peace (DH-12). Either way the mechanism is heritable wealth. Borgerhoff Mulder's coefficient for passing wealth to children is 0.13 to 0.18 among foragers and 0.61 among farmers and herders (DH-05), though another report gives 0.37 to 0.45 and I flag the range. When there is something to inherit, sons inherit it, wives are brought in from outside, and the daughters go.

Whether the incoming men were kind is another matter. When the Bell Beaker people reached Iberia about 2500 BCE, Íñigo Olalde's 2019 transect in Science found the overall ancestry shifting by about 40 per cent while the Y chromosomes were replaced almost entirely, 90 to 100 per cent, by one steppe lineage, and the same happened in Britain (DH-12). What the women's bones say is plainer. Alison Macintosh's 2017 study in Science Advances of 89 tibiae and 78 humeri from 5300 BCE to 850 CE found Neolithic women's arm bones stiffer than those of living elite rowers, from up to five hours a day at the grinding stone (DH-05). The Preface met these rowers. What it did not say is that the women's leg bones were about 12 per cent weaker by the Bronze Age than in the Neolithic (DH-04), because the work had moved indoors. At Varna on the Black Sea, in a cemetery of 4560 to 4450 BCE, the diet had already split by sex, and the richest grave, a man of 40 to 50, held more than a kilogram of gold (DH-04). Ester Boserup argued in 1970 that the plough was the turn, since it wants upper-body strength and long hours away from children (DH-05, DH-07). Then the rules got written down. The Middle Assyrian Laws of about 1076 BCE, in Law 40, set out which women must veil and which must not (DH-05).

Two more entries before I close the book. The first is on mating. Of 1,231 societies in Murdock's Ethnographic Atlas, 186 were strictly monogamous, about 15 per cent, while 453 allowed occasional polygyny and 588 practised it often (DH-09), so a man with several wives was permitted in about 85 per cent of the human record and was a reality for a few men in each. The second is the exceptions in the genomes. In Chaco Canyon, New Mexico, nine elite individuals buried between 800 and 1130 CE shared one mitochondrial genome across 330 years, a dynasty that ran through mothers (DH-12). At Çatalhöyük, the people buried together under one floor were often not kin at all (DH-12). The ledger, even after the plough, could still be kept the other way.

So here is the tool I want to hand you, because this is the second chapter of a book about the present. The entries did not vanish when we learned to write. They went under the writing. The written rules of my own lifetime, the equal pay acts and the parental leave, are amendments to a ledger that goes on being kept by nobody in every kitchen, and the older entries, about who holds the child and who is levelled and who leaves at marriage, are still being kept there, mostly by people who would say they hold no such views. My mother kept a household book in the years before I was born, and the entries that decided her days were never in it. When you next sit at a table where work is being shared out, look for the unwritten column, the one that records who is teased for wanting more, whose mother lives nearby, who is expected to move if the other one's work moves, and who does the tasks that can be interrupted. That column is the *Campfire Contract*, the bargain the bands struck without writing it, and it has not been renegotiated so much as papered over.

THE STRONGEST CASE AGAINST

The case is that the campfire was not benign for either sex, and that the ledger's oldest entries include some nobody would choose.

Start with the killing. Richard Lee, who watched the ox being insulted, also counted the dead. Among about 1,500 Ju/'hoansi between 1920 and 1970 he recorded 25 homicides, a rate of 33 to 40 per 100,000 a year, higher than the most violent American cities of the 1970s (DH-02, DH-11). About 23 per cent were over sexual jealousy (DH-02). Bruce Knauft's Gebusi of New Guinea ran at 200 to 326 per 100,000 (DH-11). Kim Hill and

Magdalena Hurtado put life expectancy at birth among the Aché at 30 to 35 years (DH-02). The children born were not all wanted equally. Juvenile sex ratios of 1.29 boys per girl among the Agta, 1.53 among the Aché and 1.72 among the Batek, reported by Abigail Page and Mark Dyble in 2019, are hard to read as anything but girls neglected or killed (DH-02). Among the Barí of Venezuela, Stephen Beckerman's 822 pregnancies show 20 to 36 per cent of children dead before 15 (DH-08).

Now the war, on which I carry three positions and rule on none. Douglas Fry and Patrik Söderberg, anthropologists then at Åbo Akademi University in Finland, took 21 mobile forager band societies and 148 lethal events in Science in 2013 and found that 55 per cent involved one killer and one victim, 85 per cent were within the killer's own society, and 10 of the 21 societies had no killing by more than one attacker at all (DH-03, DH-11). Homicide and feud, on this reading, are old. War is not. Samuel Bowles, in Science in 2009, took the higher archaeological and ethnographic estimates that Lawrence Keeley and Steven Pinker had aggregated, about 15 per cent of all deaths from violence in non-state societies, and built a model in which that death rate is exactly the pressure that selected for human self-sacrifice within groups (DH-11), with 14 per cent of deaths as the figure from his own sample. Crow Creek in South Dakota, at about 60 per cent, sits in his tail, and Brian Ferguson's objection is that Crow Creek dates to 1325 CE and is a stressed farming town, not the Pleistocene (DH-11). And at Nataruk, west of Lake Turkana in Kenya, Marta Mirazón Lahr's team reported in Nature in 2016 the remains of 27 people about 10,000 years old, of whom at least 10 had died of blows and projectile wounds, one of them a woman in late pregnancy with her hands apparently bound. Foragers, not farmers, killed together. Fry's answer would be that one site is one site, and Bowles's that one site is what the model predicts. I leave the three where they stand.

The Neolithic entries are no gentler. In Neolithic Denmark and Sweden, Linda Fibiger's 2013 study of 378 crania found healed head injuries in 16.9 per cent of men and 9.4 per cent of women, and unhealed ones equally by sex (DH-04). At Talheim 34 people were killed from behind (DH-04). At Asparn-Schletz 67 were left unburied and the young women were missing from the dead, and at Schöneck-Kilianstädten 26 were killed with their legs broken and the young women were missing again (DH-05), which is what abduction looks like in bone. The bands

that levelled their strong men did not always level them in time, and the farms that followed wrote the strong men's names on the graves.

IN PRACTICE

The ledger is still being kept, so the practical question is how to read your own copy. What follows is not a programme. It is a set of entries to look for, and what tends to happen when they are ignored.

Count the hands. If a child in your household is being raised by two adults and nobody else, you are running an arrangement for which the ledger has no entry, because no camp ever tried it. The Efe infant in 14 pairs of hands a day is the far end of the range, but the near end is not two. Sear's finding that the mother's mother is the relative who most reliably keeps children alive (DH-08) translates, in a country with vaccines, into something quieter and still real: David Waynforth's 2020 analysis of the UK Millennium Cohort found that children with a relative helping to care for them had an odds ratio of 0.61 for developmental delay (DH-03). The failure mode is familiar. A couple moves 350 kilometres for one partner's work, as the Lech valley wives did in bone, and the partner who holds the interruptible tasks finds in the second year that she is doing them alone. If the grandmother is far, the substitute is hours from the father, and the Aka benchmark is worth keeping in view, within arm's reach for 47 per cent of the day (DH-08), not a bedtime story. If neither is available, the honest entry is that the household is short of hands, and the shortage should be named as a shortage rather than as one person's failing.

Levelling needs an exit. The Ju/'hoansi could tease a hunter out of his pride because he could leave and because the teasers could be left. Take away the exit and the same mechanism turns into something else. In a family, a workplace or a friendship where one person cannot walk, because of money, a visa, a lease or a child, ridicule stops being levelling and becomes rule. If you find yourself on the giving end of the teasing, check first whether the other person could go. If you are on the receiving end and could not, it is not a campfire, whatever it is called.

Keep day talk and night talk apart. Wiessner's Ju/'hoansi complained by daylight and told stories in the dark, and conflict all but vanished after sunset (DH-03). Households that do only one fail in predictable directions. All complaint and no repair produces a ledger with debts on every

line and nobody paying. All stories and no complaint produces a household in which the entries are never read aloud, and the person holding the debit column, usually the one holding the child, finds out how large it has grown when she stops. If grievances are raised at night, in the hours the foragers kept for repair, expect them to land harder than they should.

Watch the interruptible column. Brown's rule gave women the tasks that could be broken off and resumed (DH-07), and the rule has outlived the digging stick. In a household where both adults work for money, look at which job is treated as the one that can be interrupted: the one whose holder takes the call from the school, stays home with the fever, takes the shorter contract. That job's holder is keeping the ledger, whichever sex holds it. If it is always the same person, the household has an entry it never agreed to. If you want to change the entry, the lever is not a conversation about fairness but a decision about whose work gets interrupted next time, made before the next time.

Know when you start to own things. The camp tilted when there was something to inherit, and a household tilts at the same moment, usually the first mortgage or the first business in one name. Borgerhoff Mulder's coefficients (DH-05) describe societies, not couples, but the mechanism holds at any scale: once wealth passes down a line, the line gets guarded, and the person marrying into it is a wife of Gurgy whether or not anyone means it that way. Whose name is on the deed is an entry in the ledger. So is whose parents lent the deposit. Writing rules down does not always help. When Ghana's Intestate Succession Law of 1985 gave widows and children a formal share of a husband's estate among the matrilineal Akan, the gap between men's and women's hours of work widened by 41 per cent (DH-06), because the written rule changed who could keep the unwritten ones.

Do not mistake complementarity for equality. Whyte's ten dimensions of women's standing do not move together (DH-07), and a household is no different. A couple can be equal in money and unequal in whose plans bend, equal in talk and unequal in sleep. If you audit one column and call the ledger balanced, the other columns go on being kept. The question that catches it is dull and specific: for each of the last five decisions that mattered, who proposed, who conceded, and who would have had to leave if it had gone the other way.

And do not read the average as the person. The Maasai and Khasi experiment (DH-06) shows willingness to compete reversing with the des-

cent rule, and the Mosuo inflammation figures (DH-06) show physiology following the residence rule. The person across the table is a point in a distribution that moves when the rules move. If a difference in your household looks like sex, ask first whether it is the entry, and whether the entry could be changed.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- Cooperative breeding and the maternal grandmother: Hrdy, Hawkes 1998, Sear and Mace 2008 (DH-01, DH-03, DH-08).
- Levelling by ridicule with an exit: Boehm 1993, Lee 1969, Wiessner 2014, Woodburn 1982 (DH-03).
- Division of work by interruptibility: Murdock and Provost 1973, Brown 1970 (DH-07).
- Heritable wealth as the hinge: Borgerhoff Mulder 2009, Karmin 2015 (DH-03, DH-05, DH-12).
- Patrilocality in the genomes against the mixed camp: Skov 2022, Rivollat 2023, Mittnik 2019 (DH-12) and Dyble 2015 (DH-03).

Decoration

- The Christmas ox as a scene (the rule carries, the ox does not).
- Varna's gold and the Assyrian veil (a turn already shown in bone).
- Chaco Canyon, Poro and Sande, the Gê plaza (colour, not weight).
- Wilamaya Patjxa (one burial).
- Wrangham's softened skulls (a theory told as one).

Contested

- Rates of violent death in prehistory: Fry and Söderberg 2013 vs Bowles 2009 and the Keeley/Pinker aggregate vs Nataruk (DH-03, DH-11, Nataruk not in the documents).
- Cause of the Y-chromosome collapse: Zeng, Aw and Feldman 2018 vs Guyon, Marchi, Heyer and Chaix 2024 (DH-12).
- Women hunting: Anderson 2023 vs Venkataraman and colleagues 2024 (DH-02).
- Menstrual seclusion as honesty device (Strassmann) or women's coalition (Knight) (DH-10).
- Borgerhoff Mulder's wealth-transmission coefficient: 0.37 to 0.45 (DH-03) vs 0.13 to 0.18 and 0.61 (DH-05).
- Karmin's dating: 8,000 to 4,000 years ago (DH-05) vs 5000 to 3000 BCE (DH-12).

NUMBERS TO CARRY

1. 45 populations: the maternal grandmother improves child survival almost everywhere it was measured, the father in about one third (Sear and Mace 2008, DH-08).
2. 183 of 185 societies assign cooking to women, and none assigns large-game hunting to women alone (Murdock and Provost 1973, DH-07).
3. 12 per cent of adults unrelated in a camp where both sexes choose where to live, 0.6 per cent where one sex chooses (Dyble 2015, DH-03).
4. 17 women reproducing for every man whose paternal line survived, in the millennia around 5000 to 3000 BCE (Karmin 2015, DH-05, DH-12).
5. 25 homicides among about 1,500 Ju/'hoansi over 1920 to 1970, roughly 40 per 100,000 a year (Lee 1979, DH-02, DH-11).

THE LIBRARY

- DH-01 Human evolutionary trajectory of the sexes: arm muscle, provisioning models, grandmother hypothesis, cooperative breeding.
- DH-02 Contemporary foragers, division of labour and gender relations: Hadza energetics, women hunting, leisure, homicide, life expectancy.
- DH-03 How egalitarian societies keep power in check: Boehm, Lee's ox, Wiessner's fires, Woodburn, Borgerhoff Mulder, Dyble.
- DH-04 Archaeology of gender before writing: Wilamaya Patjxa, Varna, Fibiger, Macintosh's leg bones, Talheim.
- DH-05 The Neolithic turn, agriculture, property and patriarchy: Gurgy, Karmin, Boserup, mass graves, Assyrian law.
- DH-06 Matrilineal and matrilocal societies: Minangkabau, Khasi, Mosuo, Haudenosaunee, Reynolds, Gneezy, Akan law.
- DH-07 Cross-cultural databases and what correlates with women's status: Murdock and Provost, Brown, Whyte.
- DH-08 Ancestral child-rearing and alloparenting: Ivey, Hewlett, Kramer, Meehan, Konner and Worthman, Sear and Mace, Beckerman.
- DH-09 Human mating systems and sexual selection: Murdock's polygyny counts.
- DH-10 Ritual, myth and gender complementarity: Whyte's 88 per cent, Poro and Sande, Gê moieties, Strassmann.
- DH-11 Violence, warfare and the sexes: Lee's and Knauft's rates, Fry and Söderberg, Bowles, Keeley, Pinker, Ferguson.
- DH-12 What ancient DNA reveals about the sexes: Lippold, Skov, Rivollat, Mittnik, Hazleton, Olalde, Chaco, Çatalhöyük.

CHAPTER 3

The Endocrine Accent

It is a Sunday morning in a flat where the radiator ticks, and the boy on the changing mat is six weeks old. He has done nothing yet that anyone would call male. He cries, he feeds, he sleeps in the particular heap that new babies make. His parents see a baby. His blood sees something else. Somewhere between his first and third month, the two small glands that have only just settled into place are making testosterone at a concentration his body will not reach again until he is thirteen or fourteen, a peak of 100 to 400 nanograms per decilitre that clinicians call mini-puberty and parents call nothing at all, because it shows nowhere (BI-06). It runs under the skin of an infant who cannot hold up his own head.

Across the room his mother sits in the nursing chair. Her prolactin is high. Her cycle has gone quiet, the way a river goes quiet above a weir, and her oestradiol sits low and flat (BI-02). Two bodies in one room, running two tides at different heights, and neither of them knows. By his first birthday the boy's testosterone will have sunk to almost nothing and will stay there for a decade (BI-01). In the flat downstairs a girl of the same age has follicle-stimulating hormone running high, and it stays that way until she is three or four (BI-06).

I start here because this is what a hormone is. Not an order. A weather. It is measurable in a tube of spit at nine in the morning and gone by evening. It shifts the odds of what a body does next and it decides none of it. The nearest thing I know to it is an accent. You can hear within a sentence where a voice grew up, and the accent tells you nothing about what the sentence says. That is the whole claim of this chapter, stated before

the numbers arrive so that the numbers cannot inflate it. The sexes speak with different endocrine accents. They do not speak different languages.

The claim has a shape. There is a layer where the difference is close to absolute, and it is the one you can measure in blood. There is a layer of behaviour, where the difference is shifted averages and wide overlap. And there is a layer of culture, which decides what an accent is allowed to mean. Most of the trouble in this field comes from letting a finding climb a layer it has not earned.

One more thing before the changing mat is cleared. Some of the people near the end of this chapter, who changed one sex hormone in an adult body and reported back, are not in the couple this book is about. They ran the only experiment that shows what the accent does when you change it, and I intend to listen.

Six weeks of nobody

For the first six weeks after conception a human embryo has no sex you could see. It carries one, 46,XX or 46,XY in nearly every case, but the ridge of tissue that becomes the gonads is the same ridge in both (BI-01). Around week six or seven a single gene on the Y chromosome, SRY, starts a cascade through SOX9 that turns the ridge into testes (BI-01). Without it the ridge becomes ovaries, and the old textbook phrase for that, the female default, turned out to be wrong. Ovaries need their own signals, WNT4 and RSPO1 among them, and when those fail the result is not a girl by omission but a disorder (BI-01). Nobody is the default. Both are built.

From about week eight to week twenty-four the fetal testes pour out testosterone, and a stronger cousin made from it, dihydrotestosterone, shapes the external genitalia (BI-01). Then a pause, then the surge on the changing mat, peaking at one to three months by one report (BI-06) and running to six months by another (BI-01), then a decade of near silence until puberty, which the neuroscientist Margaret McCarthy argued in Nature Neuroscience in 2015 is a second organising window and not merely a reawakening (BI-01).

That word, organising, comes from a guinea-pig laboratory. In 1959 Charles Phoenix, Robert Goy, Arnold Gerall and William Young, working at the University of Kansas, gave testosterone to pregnant guinea pigs

and watched what the daughters did as adults (AN-08, BI-01). Given ovarian hormones in adulthood, those daughters would not arch their backs for a male. Given testosterone, they mounted like their brothers. The prenatal dose had organised the tissue. The adult dose only activated what was laid down. Every argument about hormones since has been about which of those two things you are looking at.

Humans run the same experiment on themselves by accident. Girls with congenital adrenal hyperplasia, whose adrenal glands make extra androgens before birth, choose toys in a pattern shifted about 60 per cent of the way from the female towards the male average, in cohorts of 40 to 125 children (BI-01), despite parents who, when filmed, encouraged female-typical play in them more than in their sisters, in Vickie Pasterski's 2005 study of 81 such girls and 72 siblings. People with complete androgen insensitivity, 46,XY with receptors that cannot hear testosterone at all, grow up female in body and, in cohorts of 10 to 30, female in every behaviour measured (BI-01). And in 1974 the endocrinologist Julianne Imperato-McGinley reported in Science on families in the Dominican Republic where a missing enzyme left 46,XY children looking female at birth and raised as girls. At puberty, when testosterone arrived, 17 of the 18 took up male identities (BI-01). Organised early. Activated late, and heard.

One more animal result, because it corrects the mistake almost everyone makes. In 1953 Grunt and Young sorted male guinea pigs into high, medium and low copulators, castrated them all, and gave every animal the same replacement dose of testosterone (AN-08). Every animal went back to its own rank. The dose was identical. The behaviour tracked the sensitivity of the tissue, not the amount in the blood. Damassa and colleagues showed in 1977 that a tenth of the normal concentration restores a castrated rat's full sexual behaviour (AN-08). Above a threshold, more hormone does nothing. Hold on to that. It is why a man's testosterone reading tells you almost nothing about the man.

Two clocks in one kitchen

Now the adult ranges, which is where the accent becomes audible. A man's testosterone in the laboratory reference range runs from 10.4 to 34.7 nanomoles per litre, a woman's from 0.5 to 2.4 (BI-02). That is ten

to fifteen times higher and, unusually for anything in this book, the two ranges do not overlap at all. This is the first layer, the measurable one, and here it is nearly a wall. But look at what else the same blood carries. Cortisol has the same reference range in both sexes, 138 to 690 nanomoles per litre (BI-02). Oestradiol is present in men at 36.7 to 146.8 picomoles per litre, and in women it moves between 55 and 1,468 across a cycle, so that a woman in her early follicular days can sit inside the male range (BI-02). Progesterone in men is below 3.18 nanomoles per litre, in women anywhere from 0.3 to 63.6 depending on the day (BI-02). Not two chemistries. One chemistry, with one ingredient at very different levels and the others rising and falling on different timetables.

The timetables are the real difference. A man's system runs on a day. Testosterone peaks between six and eight in the morning, tied to the last stretches of REM sleep, and drifts down through the afternoon (BI-02). Every day, roughly the same shape. A woman's system runs on a month, oestradiol rising through the follicular phase, a surge of luteinising hormone at ovulation, progesterone holding the luteal weeks (BI-02). Except that the month is a fiction of the calendar. When Jonathan Bull and colleagues analysed more than 600,000 cycles logged by 124,648 app users in 2019, only 13 per cent of cycles were exactly 28 days long, the mean was 29.3, and the luteal phase averaged 12.4 days (BI-04). App users are not everyone, but a woman who says her body does not keep a 28-day schedule is, on this evidence, in the majority.

Stress runs through the same plumbing and shows a second, smaller difference. In the Trier Social Stress Test, a mock job interview in front of a stony panel, men on average mount a higher cortisol peak than women, across 34 studies with 640 women and 710 men in a 2017 meta-analysis (BI-02), while women report more distress, a pattern Brigitte Kudielka and Clemens Kirschbaum described in Biological Psychology in 2005 (BI-06). More hormone and less reported distress on one side, the reverse on the other. Neither is the true measure of the stress. Both are.

Sleep is where I would keep a hand on the rail. The finding everyone repeats, from Rachel Leproult and Eve Van Cauter in JAMA in 2011, is that a week of five-hour nights cut young men's testosterone by 10 to 15 per cent (BI-02, BI-05). It was ten men. A 2021 meta-analysis of 18 controlled studies in Sleep Medicine found that partial restriction of that kind produced a standardised mean difference of −0.22 that did not reach signific-

ance, and that only total sleep deprivation reliably lowered testosterone, by about a quarter (BI-05). One report calls the ten-man result robust and another calls it fallen. I call it contested, and small.

Cebu, 2005 to 2009

In 2005 the anthropologist Lee Gettler began collecting the saliva of young single men enrolled in a long-running birth-cohort study in Cebu, in the Philippines (BI-02, BI-06). He collected it again four and a half years later. The men who had started with the highest testosterone were the most likely to have found a partner and become fathers in between (BI-02). And the new fathers had lower testosterone than the men who had stayed single. The median fell by 26 per cent on waking and 34 per cent in the evening according to three of my reports, and by 20.5 per cent according to a fourth (BI-02). The men doing three or more hours of childcare a day had the lowest readings of all (BI-06). The sample is given as 394, 465, 624 and 683 men in different summaries, which is a fair reminder that I am reading summaries and not the paper. What does not vary between them is the direction. Fatherhood, and specifically the hands-on kind, lowers a man's testosterone. A 2019 meta-analysis by Nicholas Grebe and colleagues pooling 114 effects from 66 studies confirmed the direction, partnered men below single and fathers below childless, with effect sizes my extraction lost (BI-05). The Campfire Contract of chapter 2 rests on this.

The hormone is not only answering nappies. It answers rivals, winning, and a stranger with a nice face. When the psychologist Shawn Geniole and colleagues pooled 60 effects from more than 2,500 participants in Hormones and Behavior in 2017, winners of contests showed a larger rise in testosterone than losers, and the effect was larger on the tennis court and the judo mat than in a laboratory pressing buttons (BI-05, effect size stripped,). This is the challenge hypothesis, borrowed from the field biologist John Wingfield's seasonally breeding birds in the American Naturalist in 1990 (AN-08), and its 2019 revision by Wolfgang Goymann and colleagues in BioScience adds the correction most popular tellings miss. Across vertebrates, courtship and mate-guarding provoke bigger testosterone pulses than fights do (AN-08). It is a hormone of reproduction and

standing before it is a hormone of aggression.

Then the question everyone actually wants answered. Does it make men aggressive? Geniole's 2020 meta-analysis, 45 studies and 9,760 people, put the correlation between baseline testosterone and aggression at r = 0.054, with a confidence interval from 0.028 to 0.080 (BI-03, AN-08). One report gives the overall figure as 0.14 instead (BI-02). Take the larger number and it accounts for 2 per cent of the variance. Take the smaller and it accounts for about a third of one per cent. When testosterone was given rather than measured, the correlation was 0.046 and not significant (BI-03). Remember Christoph Eisenegger's women in the Preface, moved more by the belief than by the hormone. In 2024 the economist Amos Nadler with Zava and Camerer gave 1,000 men a single 11 milligram intranasal dose and found no effect on trust, risk, competitiveness or fairness (BI-03), although one of my reports attributes the same 1,000-man trial to Gideon Nave and colleagues in 2018. What the hormone does seem to do, in the psychologist Justin Carré's 2017 study of 121 men in Biological Psychiatry, is raise aggression only in men who were already dominant and low in self-control (BI-03). In Jean-Claude Dreher's 2016 study of 40 men it raised both punishment of unfair offers and reward of generous ones (BI-12). The accent got thicker. The sentences stayed the speaker's own.

Levels themselves are also a place. Tsimane men in the Bolivian Amazon carry about 33 per cent less testosterone than age-matched Americans and show no decline with age, in the anthropologists Benjamin Trumble and Michael Gurven's studies, yet 63 of them showed a 30 per cent surge after a football match (BI-10). Low baseline, full response. The high and falling testosterone of the Western man may be a mark of surplus calories and few parasites rather than of maleness itself (BI-10).

Thirteen per cent

Now the other column, and a hut. Between 1986 and 1988 the anthropologist Beverly Strassmann kept a 736-night census of who slept in the menstrual hut of Sangui, a Dogon village in Mali (BI-04). Dogon women were obliged to spend their periods there, and her count showed how rarely that happened. With late menarche, many pregnancies and long

breastfeeding, a Dogon woman had something like 110 menses in her life, against 400 to 450 for a woman in a modern city (BI-04). Another report puts the ancestral figure at 100 to 160 (BI-10). The monthly cycle as a monthly experience is a modern arrangement. Dogon men, Strassmann found, watched the hut. It told them who could be pregnant, and by whom (BI-04).

What the cycle does to the mind is a smaller matter than the culture around it suggests, and the honest way to say that is to list what fell. The finding that women prefer more masculine faces and more dominant men around ovulation rested on a 2014 meta-analysis by Kelly Gildersleeve and colleagues, 134 effects from 50 studies, that was contradicted the same year by the psychologist Wendy Wood's re-analysis of 58 reports (BI-04). Then came larger, better-built studies. Benedict Jones and colleagues in Psychological Science in 2018, 584 women and more than 5,000 test sessions, found no link between measured hormones and face preference (BI-04). Julia Stern's 2021 study of 257 women and 1,028 sessions found the same (BI-04). What survived is duller and real. Oestradiol tracks desire upward and progesterone tracks it downward, in James Roney and Zachary Simmons's 2013 work, and across 26,000 diary entries Ruben Arslan's 2021 study found a mid-cycle peak in desire (BI-04). Partners cannot tell, in Lara Schleifenbaum's 2022 study (BI-04). The tide moves. Nobody on the shore sees it.

Then the minority for whom the cycle is not an accent but a front coming through. Pooled across 17 studies, 47.8 per cent of women report premenstrual symptoms (BI-04). Between 3 and 8 per cent meet criteria for premenstrual dysphoric disorder (BI-04), or about 5 per cent by another report (BI-03). In 1998 Peter Schmidt and David Rubinow at the National Institute of Mental Health in Bethesda quietened the ovarian cycle of 22 women with leuprolide and then added back oestradiol and progesterone under blinding, reporting in the New England Journal of Medicine (BI-03, BI-04). Women with the disorder got their symptoms back from the add-back. Women without did not. The hormones were the same. The response was not, which is Grunt and Young's guinea pigs again in a different key. A 2017 follow-up in the American Journal of Psychiatry found the symptoms return in the first month of add-back and fade by the third, so it is the change and not the level that hurts (BI-04). At the scale of a country, a 2019 survey of 32,748 Dutch women in BMJ

Open put the cost at 8.9 days of lost productivity a year, with 80.7 per cent working while unwell and 20 per cent having told an employer why (BI-04).

One in two hundred

The most common way to change a woman's hormones is a small tablet. In 2016 Charlotte Wessel Skovlund and colleagues in Copenhagen followed 1,061,997 Danish women aged 15 to 34 through the national registers and found that starting hormonal contraception raised the chance of a first antidepressant prescription, a relative risk of 1.23 for the combined pill, 1.34 for the progestogen-only pill, 1.40 for the hormonal coil and 1.60 for the vaginal ring, with the highest figures in adolescents (BI-08, BI-10). In absolute terms the rate went from 1.7 to 2.2 per 100 woman-years, one extra woman in 200 (BI-10). The same group reported in 2018 on 475,802 women, 6,999 suicide attempts and 71 suicides, with a relative risk of 1.97 for attempts and 3.08 for suicide, peaking two months after starting (BI-08). A 2023 UK Biobank analysis of 264,557 women, with 7,354 sibling pairs to hold family background still, found a hazard ratio of 1.71 for depression in the first two years of use (BI-08). The randomised trials are small and say something softer. Niklas Zethraeus and colleagues in 2017, 340 women, found lower general well-being on the pill and no difference on a depression scale (BI-08). Registers say risk. Trials say mood. Most women in every dataset say nothing changed, and they are the majority.

The plausible mechanism is a missing neurosteroid. The body's own progesterone is broken down into allopregnanolone, which calms the same GABA-A receptors a sedative does, and synthetic progestins are not (BI-10). The pill also raises sex-hormone-binding globulin by 80 to 300 per cent, which mops up free testosterone, and about 15 per cent of users report lower desire while more than 80 per cent report no change, in a 2013 review by Pastor and colleagues (BI-08). Whether the pill changes whom a woman chooses is where I stop trusting. The congruency hypothesis, from Craig Roberts's 2014 study of 365 couples, held that women who met their partner on the pill and later came off it lost attraction (BI-10). One of my reports calls this replicated (BI-10). Another says Julia Stern's

attempts in 2019 and 2021 failed to replicate it (BI-08). Two reports, opposite verdicts. I file it under contested and carry on.

The thermostat and the duvet

What matters in a shared flat is rarely the hormone. It is what the hormone has built. Start with the clock. In 2011 the sleep researcher Jeanne Duffy and colleagues in Boston measured the intrinsic circadian period of 157 people, 52 women and 105 men, away from daylight, and found 24.09 hours in women and 24.19 in men (BI-09). Six minutes. But 35 per cent of the women ran shorter than 24 hours against 14 per cent of the men, which tilts women earlier (BI-09). Women get more slow-wave sleep and better sleep by the instruments across 169 studies in a 2019 review in Sleep, and report insomnia 1.41 times as often, while men have more apnoea (BI-09). She sleeps better and says she sleeps worse. He snores. Both are measured.

Then the thermostat. In 2015 Boris Kingma and Wouter van Marken Lichtenbelt at Maastricht showed in Nature Climate Change that the standard office comfort model, ASHRAE 55, had been calibrated on a 70-kilogram man's metabolism and could overestimate women's heat production by up to 35 per cent (BI-09). Their sample was 16 women. Bjarne Olesen, who helped write the standard, replied that clothing insulation explained most of the gap (BI-09). Women in the studies prefer 24 to 25 degrees and men 21 to 22, and the ratio of women to men dissatisfied with office temperature was 1.74 in a 2011 review by Karjalainen (BI-09). The luteal phase adds 0.3 to 0.5 degrees to a woman's set-point (BI-09). None of this is culture. All of it is a matter for negotiation over a duvet.

Pain runs the other way. Women have lower pain thresholds and lower tolerance, d = 0.66 and d = 0.79 in reviews by Riley in 1998 and Racine in 2012, distributions that still overlap by roughly three-quarters (BI-09). And men's thresholds drop when the experimenter is a woman (BI-09). The physiology is real and the third layer walks in wearing a lab coat. Muscle is where the accent is loudest. Men carry about 60 per cent more, most of it above the waist, in William Lassek and Steven Gaulin's 2009 figures, and women about 1.6 times the body fat (Preface, SY-01). The Preface gave you the grip-strength gap from Dieter Leyk's 2,187 Germans,

and I leave it there. It is the one place where the sexes barely overlap, and the one that matters least at a kitchen table.

The room the accent is spoken in

The levels are moving. Soum Lokeshwar and colleagues reported in European Urology Focus from the American NHANES survey of 4,207 men aged 15 to 39 that mean testosterone fell from 605 to 451 nanograms per decilitre between 1999 and 2016, and from 665 to 529 among men of normal weight (BI-05, BI-10). Sperm concentration fell 51.6 per cent between 1973 and 2018 across 223 studies in 53 countries in Hagai Levine's 2022 meta-regression, and the fall ran at 2.64 per cent a year after 2000 (BI-10). This last is fought over. Marion Boulicault and colleagues at Harvard argued in 2021 that counts vary across a wide range without being pathological, and Allan Pacey at Sheffield that the laboratory methods changed (BI-10). Girls' breast development advances by 0.24 years a decade in Camilla Eckert-Lind's 2020 meta-analysis of 30 studies (BI-10). Phthalates from plastics shorten the anogenital distance of newborn boys, a marker of how much androgen reached them before birth (BI-10). I carry all this as the room and not the speech. It is an accent spoken in a house where somebody keeps moving the furniture.

People who reported back

Here I have to say something plainly. The couple at the centre of this book is a man and a woman. The people in this section are, for the most part, not in that couple. They are transgender men and women in hormone therapy, men with prostate cancer whose testosterone was medically removed, and, earlier, the intersex children of the Dominican families. They appear because they have done what no ethics committee would let me do to a volunteer, which is to change one sex hormone in an adult body and describe it from the inside. Their reports are evidence, not a footnote.

What they say is more consistent than I expected. In Ghent, Justine Defreyne and the ENIGI consortium followed 898 people, 440 trans men and 468 trans women, through 36 months of hormone therapy and found

no correlation between serum testosterone and state anger at 3, 12 or 36 months (BI-12). Trans men's anger scores rose slightly at three months and were back at baseline by a year, and the people who were angry were those who had arrived with depression and interpersonal sensitivity (BI-12). Daryl O'Connor, John Archer and Frederick Wu had found the same shape in 28 cisgender men given 1,000 milligrams of testosterone undecanoate in 2004. Subjective anger rose from 7.48 to 10.71 on the mood scale, and aggressive behaviour did not move (BI-12). The feeling changes. The act does not follow.

What does change is smaller and stranger. In Tim van de Grift's interviews with 26 people, 12 on feminising hormones and 14 on testosterone, those on oestrogen described crying more often and more easily, and those on testosterone described the opposite, some of them a physical block behind the eyes when grief was present and the tears were not (BI-12). Men on androgen deprivation for prostate cancer describe the same lowered threshold from the other side, and many of them hate it (BI-12). In the ENIGI cohort, 73.9 per cent of trans men reported a marked rise in sexual desire soon after starting testosterone, and it was solitary desire that rose, not desire for a partner (BI-12). Brains change measurably. In Hilleke Hulshoff Pol's 2006 study, testosterone raised trans men's total brain volume by an average of 31 millilitres in four months (BI-12). What does not change reliably is cognition. Stephanie van Goozen's 1995 finding that testosterone raised spatial scores and lowered verbal fluency was only half replicated in 1999, the spatial half (BI-12), and Sarah Burke's 2016 imaging study found people performing like their lived gender before any hormone had been given (BI-12). Testosterone as an antidepressant works in a 2019 meta-analysis of 27 trials and 1,890 men and fails in a 2023 pooled analysis of 5,601 men with low levels (BI-12). Small numbers, large expectations, as the report itself warns. But when the direction of a change is the same in Ghent, in Utrecht, in a prostate clinic and in a guinea-pig cage, I stop calling it anecdote.

So this is what the memoirists tell me, at my own kitchen table at the hour when my own testosterone is highest and about to start down. Change the hormone and the threshold for tears moves. Desire moves. The sense of a stable sky moves. The person does not. Anger does not become violence, politics do not change, the same jokes are still funny. The accent is thick enough to hear from across the room, and it decides

nothing that is said. Which leaves the question this chapter cannot answer alone: if the accents differ this much and the sentences this little, where do the differences in behaviour that do exist come from? That is the next chapter, and the Overlapping Bells.

THE STRONGEST CASE AGAINST

The strongest case against this chapter is not that hormones are unreal. It is that they are so weak a predictor of anything a couple argues about that a chapter on them is a long way of saying not much. Take the number I have leaned on most. Geniole and colleagues in 2020 found that baseline testosterone and aggression correlate at r = 0.054 across 9,760 people (BI-03, AN-08). If a friend told you two things were correlated at 0.05, you would assume she was making a point about correlation. Nadler and colleagues gave 1,000 men the hormone in 2024 and moved nothing (BI-03). The ENIGI cohort followed 898 people through hormone therapy and found no relation between the hormone and anger (BI-12). If dosing a thousand men and following nine hundred people through transition moves nothing you would put in a book, then testosterone is not an accent. It is a rumour.

The second prong is that what does move, moves with culture. The Preface showed a border shifting mothers' working hours by 350 a year without touching a single hormone. The pill's absolute effect on depression is one woman in 200 (BI-10). The cycle effects that made headlines fell when the samples got large (BI-04). Eisenegger's women changed their behaviour on belief alone. Cordelia Fine in Delusions of Gender and Rebecca Jordan-Young in Brain Storm argue that the organising-activating model imports a rodent determinism into a species that Kim Wallen, writing in Hormones and Behavior in 1999, showed had been emancipated from it, in that a human can mate, fight or care at any point in the hormonal cycle and a female rat cannot (AN-08, BI-12). If the cortex can veto the hormone, the hormone is a suggestion, and a chapter about suggestions should be short.

The third prong is measurement. Oxytocin assays disagree threefold in the same blood (BI-02). My own reports disagree on Gettler's sample size by nearly a factor of two and on Geniole's correlation by a factor of

three. A field that cannot agree on its own numbers should not be asked to explain a marriage.

My reply is short, because the case is partly right. The ranges of testosterone in the blood do not overlap, and no cultural account has ever made them. The direction of change in the people who altered their hormones is the same across cohorts, methods and countries, which rumours do not manage. And the weak correlations are the point of the chapter, not an objection to it. An accent that decided sentences would be a language. I have said it is not. The critics and I disagree about the size of the type in which that should be printed.

IN PRACTICE

What follows is conditional, because that is all the evidence licenses.

If you are a man and you notice you are sharpest and quickest to a fight before nine in the morning, that may be the daily peak (BI-02). The failure mode is using it as a reason. The correlation between the hormone and aggression is 0.05, which leaves you owning the other 99.7 per cent of the variance.

If your partner's cycle is not 28 days, she is in the majority, by the 13 per cent figure from 600,000 logged cycles (BI-04). The failure mode is a calendar app that assumes otherwise and a partner who believes the app over the woman.

If a woman reports premenstrual symptoms, the odds are close to even that she is one of the 47.8 per cent who do (BI-04), and if the symptoms are severe enough to change the household, she may be in the 3 to 8 per cent for whom the change in hormone level, and not the level, is the trouble (BI-04). The failure mode on one side is a man who hears an excuse. The failure mode on the other is a woman who does not tell a doctor, because this one has a treatment.

If she started or stopped hormonal contraception in the last few months and her mood has moved, count the months before you count the arguments. The register data peak at two to six months after starting (BI-08, BI-10). The failure mode is attributing everything to the pill, when the absolute risk is one woman in 200, and its mirror, attributing nothing to it because most women are fine.

If you argue about whether the room is at 21 or 24 degrees, both of you are likely to be physiologically right, and the office standard was calibrated on one of you (BI-09). The failure mode is turning a set-point into a character trait. Always cold is a temperature, not a personality.

If she says she slept badly and the tracker says she slept well, both are probably true, because women sleep better by the instruments and report worse (BI-09). The failure mode is the tracker winning the argument.

If, under the same pressure, he goes quiet and she says how bad it is, the Trier data suggest that his cortisol may be the higher of the two and her reported distress the higher (BI-02, BI-06). The failure mode is each reading the other's display as the whole of the other's state. Silence is not calm, and saying it is not weakness.

If one of you has just won or lost something, a match, a pitch, a row with a sibling, expect a short tide in testosterone that is larger for real contests than for games, by the 2017 pooling of 60 effects (BI-05). The failure mode is scheduling the difficult conversation for the hour after the final whistle.

If a new father finds his drive lower and his patience longer, that is likely to be Cebu (BI-02, BI-06). The drop is largest in men doing three or more hours of childcare a day, so it tends not to happen to fathers who are never in the room. The failure mode is reading it as a loss of manhood, or expecting the softening without doing the hours that produce it.

If a man has a testosterone reading that the laboratory calls low and he feels fine, remember the guinea pigs. Above a threshold, more hormone changes nothing (AN-08), and correcting a marginal deficit did not lift mood in 5,601 men (BI-12). The failure mode is treating the reading as the man, and paying for the treatment.

If someone in your household is changing their hormones, through menopause, through cancer treatment or through transition, expect the thresholds for tears and desire to move and the person to stay (BI-12). The failure mode is grieving a person who is still there.

If you are both past fifty, expect the accents to converge. Her oestradiol has fallen and his testosterone has been drifting down at roughly 1 per cent a year since thirty (BI-06). The failure mode is a script written for the bodies you had at twenty-five, and the couples who do best at this age seem to be the ones who have stopped reading from it.

If you are trying to conceive and reading the headlines about sperm

counts, know that the 51.6 per cent decline is a real number from 223 studies and also a contested one, with serious critics at Harvard and Sheffield (BI-10). The failure mode is treating a population trend as a personal diagnosis. A clinic can measure the individual. A headline cannot.

If you are tempted, anywhere in this chapter, to say that men are one way and women another, say it and then say the overlap in the same breath. Every behavioural difference here sits on distributions that mostly coincide, and every physiological one sits under a culture that decides what it may mean. That habit is the whole practice. The rest is furniture.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing: the non-overlapping testosterone ranges (BI-02). The organising-activating distinction and its human natural experiments (AN-08, BI-01). Threshold not dose, from Grunt and Young and from Damassa (AN-08). The direction of the fatherhood drop in Cebu (BI-02, BI-06). The consistency of first-person reports across ENIGI, O'Connor and van de Grift (BI-12). The register direction on contraception and mood, with its small absolute size (BI-08, BI-10).

Decoration: the six-minute clock (BI-09). The Dogon hut (BI-04). The Tsimane football match (BI-10). The morning peak at my own kitchen table.

Contested: sleep restriction and testosterone, Leproult against the 2021 meta-analysis (BI-02, BI-05). The sperm decline, Levine against Boulicault and Pacey (BI-10). The congruency hypothesis, Roberts against Stern (BI-08, BI-10). The thermostat, Kingma against Olesen (BI-09). Causation in the pill data, registers against trials (BI-08). Testosterone for depression, the 2019 meta-analysis against the 2023 pooled analysis (BI-12). Fallen and named as such: the ovulatory face preference, Jones 2018 and Stern 2021 (BI-04). Testosterone and mind-reading, Nadler 2019 (BI-12).

NUMBERS TO CARRY

1. Testosterone reference ranges: men 10.4 to 34.7 nanomoles per litre, women 0.5 to 2.4, no overlap (BI-02).
2. Baseline testosterone and aggression: r = 0.054, confidence interval 0.028 to 0.080, 45 studies, 9,760 people, Geniole et al. 2020 (BI-03, AN-08). One report gives 0.14 (BI-02).
3. Fatherhood in Cebu: median testosterone down 26 per cent waking and 34 per cent evening, or 20.5 per cent in one report, over 4.5 years, sample given as 394 to 683, Gettler et al. 2011 (BI-02, BI-06).
4. Hormonal contraception and antidepressant use: 1.7 to 2.2 per 100 woman-years, relative risk 1.23 to 1.60 by method, 1,061,997 women, Skovlund et al. 2016 (BI-08, BI-10).

5. Cycle length: 13 per cent of more than 600,000 logged cycles exactly 28 days, mean 29.3, Bull et al. 2019 (BI-04).

THE LIBRARY

BI-01: how a body becomes male or female, the natural experiments, and the organising-activating frame.

BI-02: the two endocrine systems compared, reference ranges, daily and monthly rhythms, stress.

BI-03: hormones and behaviour, Geniole 2020, Carré, Nadler, oxytocin.

BI-04: the menstrual cycle and mind, Strassmann, PMDD, what fell in replication.

BI-05: male hormonal dynamics, winner effect, sleep, Grebe 2019.

BI-06: the life course from mini-puberty to old age, TSST, Gettler childcare hours.

BI-08: hormonal contraception, Skovlund, Zethraeus, congruency.

BI-09: bodily differences in a shared flat, clock, thermostat, pain, immunity.

BI-10: the modern endocrine environment, Levine, Travison, Tsimane, thelarche.

BI-12: first-person evidence from hormone therapy, ENIGI, O'Connor, van de Grift, Hulshoff Pol.

AN-08: hormones and behaviour in animals, Phoenix 1959, Grunt and Young, challenge hypothesis, Wallen.

CHAPTER 4

Overlapping Bells

She is standing in a room without a window, four minutes into a speech she was given ten minutes to prepare, and the three people at the table in front of her have not nodded once. A camera light is on behind them. She has been told to apply for a job she does not want, to a panel trained to show nothing, and in a moment they are going to ask her to count backwards in steps of seventeen and to start again whenever she slips. Under her tongue a roll of cotton is soaking up the only thing in the room that anyone is really interested in.

This is the Trier Social Stress Test, the standard laboratory stressor, and the cotton is for cortisol, the hormone the adrenal glands release when the body decides that something matters. In seven or eight people out of ten the cortisol in that roll rises to between two and four times its resting level within about twenty minutes, according to the PS-04 summary of the protocol. That is the physiological layer, and it is measurable to the nanomole. What the cotton cannot tell you is whether the person holding it is a man or a woman. The average man's cortisol rise is one and a half to two times the average woman's, a figure Kudielka and Kirschbaum, the stress researchers whose 2005 review in Biological Psychology pooled the Trier results, put at the centre of their account. Yet the woman in the room may be in the luteal half of her cycle, in which case her curve matches the men's. She may be on the pill, in which case it is blunted. She may simply be one of the many women whose adrenal glands answer a panel of blank faces harder than most men's do.

Draw it. Take a sheet of paper and a pencil, put cortisol along the bot-

tom, and stack every man's peak into one hump and every woman's into another. You get two bells on one axis, one set a little to the right of the other, and most of their area shared. The tips of the bells are at different places. The bulk of each bell sits on top of the other. Every honest claim about the sexes in this chapter has that shape, and the whole argument comes down to two questions you can ask of any pair of bells: how far apart are the tips, and how much of the area is shared.

I want to give you that drawing as a tool, and then I want to show you what happens to it when you stop asking people about themselves and start measuring them, when you change the room, and when you stop looking at one bell at a time and look at fifteen at once. The picture does not stay simple. But it stays a picture of two bells, and you can carry it in your pocket.

The shape of a small difference

The Preface told you what Hyde found in 2005 and what Zell and colleagues found ten years later: that most psychological sex differences are small, and that the average one puts the two bells about a fifth of a standard deviation apart. I am not going to retell that. I am going to do the arithmetic underneath it, because the arithmetic is the whole tool.

A difference between two bells is written as d, the distance between the tips in units of the bells' width. The BI-11 report gives the conversion table, and the table is worth more than any adjective. At $d = 0.20$, the difference Zell's average sits near, the two bells share about 92 per cent of their area, and if you drew one man and one woman at random the person from the higher group would score higher 56 times in a hundred. At $d = 0.50$, the shared area is about 80 per cent and the random draw goes the expected way 64 times in a hundred. At $d = 0.80$, a difference psychologists call large, the shared area is still 69 per cent and the draw goes the expected way 71 times in a hundred. At $d = 1.00$ the shared area is 62 per cent. For comparison, the sex difference in human height is about $d = 2.6$, a figure PS-05 takes from Nelson 2012, and at that distance the bells share 19 per cent and the random draw goes the expected way 96 times in a hundred. Height is what a categorical difference looks like. Almost nothing in the mind looks like height.

Now put real traits on the axis. Janet Shibley Hyde, the psychologist at the University of Wisconsin who built the similarities hypothesis, has spent a career counting, and her name is on half the meta-analyses in this chapter, which is why I have chosen her as the person to follow through it. Personality first. Costa, Terracciano and McCrae analysed 26 cultures and 23,031 people in 2001 and found women higher on neuroticism, $d = -0.51$ in the American sample, and on agreeableness, $d = -0.59$, in the convention where a negative sign means women score higher. Schmitt, Realo, Voracek and Allik repeated the exercise across 55 nations and 17,637 people in 2008 and found the same two traits, neuroticism at $d = -0.40$ worldwide and agreeableness at −0.15. Kajonius and Johnson used an American public sample of 320,128 in 2018 and got −0.40 and −0.58. Those are behavioural averages from self-report, and at that size the bells share around 80 per cent of their area. The layer is behaviour, not physiology, because a questionnaire measures what people say about themselves, and what people say is a cultural act.

The picture sharpens when the broad traits are broken into their facets. Within extraversion, women score higher on warmth and gregariousness and men on assertiveness and excitement-seeking, and the two cancel when you add them up, so the domain score reports no difference where the facets report two. A model built to keep such things apart, the six-factor HEXACO, finds about one full standard deviation on the factor it calls emotionality, women higher, according to PS-01, which names no primary author for the figure. That is a big bell-shift, and at that distance the bells still share 62 per cent of their area.

The gap that is not small is interests. Su, Rounds and Armstrong pooled 47 interest inventories and 503,188 respondents in 2009, and the Preface has already told you the headline, $d = 0.93$ on the axis that runs from working with things to working with people, men towards things. What the Preface did not give is the shape underneath. The realistic scale, tools and machines and outdoor work, sits at $d = 0.84$. The social scale, teaching and helping and nursing, sits at −0.68. The enterprising scale, selling and leading, sits at 0.04, which is nothing. The bells are far apart on one axis and on top of each other on the one next to it, and the axis where they sit apart is the axis that predicts what people choose to do for forty years. That is behaviour, and the cultural layer has a hand on it, but it is behaviour that shows up in the same direction in every country that

has been surveyed, which is one of the things the strongest case against has to explain.

Two more axes belong here because they matter for what comes later. Else-Quest, Hyde, Goldsmith and Van Hulle pooled 35 dimensions of temperament in children from three months to thirteen years in 2006. Girls led on inhibitory control at $d = -0.41$. Boys led on activity level at 0.33. On negative affectivity, the tendency to sadness and fear, there was no difference to speak of. Whatever produces the adult gap in anxiety and depression, it is not there at three months. And Cross, Copping and Campbell, pooling 741 effect sizes in 2011, found no sex difference in reward sensitivity, $d = 0.01$, and a female lead in sensitivity to punishment at −0.33. Men do not want the prize more. They mind the penalty less.

Then there is the one everybody has heard of. The Preface put mental rotation at about $d = 0.57$. Voyer, Voyer and Bryden's 1995 meta-analysis in the Psychological Bulletin, which BI-11 treats as the standard, gives 0.73 for mental rotation, and BI-07 reports values ranging up to 1.0 depending on the task. The range from 0.57 to 1.0 across reports is not resolved here. At 0.73 the random draw goes the man's way about seven times in ten, and roughly 23 per cent of women score above the median man. That is a large difference by the standards of the mind, and it still leaves nearly a quarter of women above half the men.

Where the gap goes when you stop asking

Here is the finding that organises the middle of this chapter, and I want you to watch for it everywhere from now on. Whenever a sex difference is measured by asking, it is larger than when the same difference is measured by looking, and it is larger again than when it is measured by instrument. The bells drift apart as you climb from physiology to behaviour to culture, and much of what the culture calls a difference in men and women is a difference in what men and women say.

Emotion is the clearest case. Chaplin and Aldao pooled 555 effect sizes from 166 studies of children and adolescents, 21,709 of them, in 2013. Girls expressed more positive emotion at $g = -0.08$ and more sadness and anxiety at −0.10, while boys expressed slightly more anger at 0.09, figures so small that the bells are all but one bell. The differences shrank to nothing

when children were watched alone or with their parents and grew when they were watched with unfamiliar adults or in peer groups of their own sex. A child does not change its endocrine system when a stranger walks in. It changes its display. Eisenberg and Lennon showed in 1983 that women's lead in empathy on questionnaires vanishes under skin conductance and facial electromyography, and Klein and Hodges showed in 2001 that men's shortfall in reading another person's feelings vanishes when there is money on the table. Kirkland and colleagues pooled 42 samples and 4,290 people on the Reading the Mind in the Eyes test in 2013 and found a female lead of $g = 0.177$, and a 2022 study of 312,739 people across 57 countries in the Proceedings of the National Academy of Sciences found the same direction everywhere. That is a real behavioural difference, cross-cultural and stable. It is also about a sixth of a standard deviation, which means the bells share more than 90 per cent of their area, and it moves when the incentives move.

What does not move under measurement is the pattern of what people do with the feeling once they have it. A 2023 meta-analysis of 120 studies and 88,721 people found men score higher on alexithymia, the difficulty of naming one's own feelings, at $d = 0.22$ overall and 0.49 on the sub-scale for externally oriented thinking. None of these figures is a difference in feeling. They are differences in what the room lets you do with it.

The same drift shows in talk. The claim that women speak 20,000 words a day and men 7,000 went around the world in the 2000s without a study behind it. Mehl, Vazire, Ramírez-Esparza, Slatcher and Pennebaker put a recorder on 396 students, published in Science in 2007, and let it capture thirty seconds of sound every twelve and a half minutes for days at a time. The women averaged 16,215 words a day and the men 15,669, with the range in both sexes running from about 500 to about 47,000. The honest meta-analyses find small differences in the other direction. Leaper and Ayres pooled 73 studies in 2007 and found men talk slightly more, $d = -0.14$, and Anderson and Leaper pooled 43 studies of interruption in 1998 and found 0.33 for intrusive interruptions, men doing more of them, against the 96 per cent that Zimmerman and West had counted in 1975 in a sample small enough to be an anecdote. Every one of those bells shares more than 85 per cent of its area with the other.

And one of them, the biggest, dissolves when you change the seating. LaFrance, Hecht and Paluck pooled 448 effects from 162 reports of smil-

ing in 2003 and found women smile more at $d = 0.41$, a moderate gap. Then they sorted by role. When men and women were in the same social role, the same job at the same rank, the gap went away. The smile was contingent. It was the trace of a cultural position, not of a sex.

Beneath the talk, the brain gives you nothing to hold onto. Sommer, Aleman, Bouma and Kahn pooled 14 imaging studies of language lateralisation in 2004, 377 men and 442 women, and got $d = 0.21$ with a confidence interval from −0.05 to 0.48, which crosses zero. The physiological layer is silent here. The gaps in talking are behavioural and cultural, and they are small.

One more figure belongs here because it is the one couples' therapists quote most. Gottman found that when a partner's heart rate passes 100 beats a minute in an argument, roughly, the person floods, the thinking part of the conversation stops, and recovery takes twenty to thirty minutes. That is physiology and it is robust. The figure that usually follows it, that 85 per cent of stonewallers are men, is sourced in both PS-02 and PS-03 to a couples' app blog and not to Gottman's own data, so I give it here as a claim and not as a fact. What the peer-reviewed literature does give is Schrodt, Witt and Shimkowski's 2014 pooling of 74 studies and 14,255 people on the demand-withdraw pattern: women demand and men withdraw on average, and the pattern reverses when it is the husband who wants the change. The withdrawal belongs to whoever is being asked to move, not to a sex.

Cortisol and the hand on the shoulder

Back to the cotton roll. The Trier results are the one place in this chapter where the physiological layer speaks clearly, and it says three things. First, on average men's cortisol answer to a hostile panel is bigger than women's, by the factor of 1.5 to 2 that Kudielka and Kirschbaum reported. Second, Kirschbaum and colleagues showed in 1999 that the bell for women is really several bells drawn at different points of a month. Third, the difference is in the kind of threat. Stroud, Salovey and Epel gave fifty people two stressors in 2002, an achievement task and a social rejection, and found the men's cortisol rose to the achievement task and the women's to the rejection. Same gland, different trigger.

Now the finding I would keep if I could keep only one from this chapter. Kirschbaum, Klauer, Filipp and Hellhammer, in Trier in 1995, ran the stress test with the subject's partner in the room, instructed to give verbal support before the ordeal. The men whose partner spoke to them produced less cortisol. The women whose partner spoke to them produced, on average, slightly more. Ditzen and colleagues followed it up in 2007 with 67 women: a ten-minute neck-and-shoulder massage from the partner before the test lowered cortisol and heart rate, and spoken support did not. The partner-support paradox, then, is this: the average man is calmed by being told it will be fine, and the average woman is calmed by being held and not by being told. Ditzen's 2009 trial in Biological Psychiatry, 47 couples, 94 people, gave intranasal oxytocin, the hormone of the milk let-down and the embrace, before a scripted argument and got more eye contact and less cortisol in both sexes, but her 2013 follow-up found the drug lowered salivary alpha-amylase, the sympathetic marker, in women and raised it in men. Same molecule, opposite direction. This is all physiology, all small samples, and all worth checking. The point for a household is not that the pattern holds in your kitchen. The point is that the two ways of helping are not interchangeable, and that people reach for the one that would work on them.

The cost of the pattern shows up in the clinic. Salk, Hyde and Abramson, that Hyde again, pooled more than 3.5 million people across 90 nations in 2017. Women are diagnosed with depression at an odds ratio of 1.95, and in symptom terms $d = 0.27$, a small bell-shift with more than 85 per cent shared area. The gap opens at twelve, when the odds ratio is 2.37, peaks between thirteen and fifteen at 3.02, and then narrows, and it is larger in nations with more gender equality, not less. Martin, Neighbors and Griffith asked in 2013 whether the diagnosis is measuring the disease or the display. In 5,692 Americans, the standard symptoms gave the usual female lead. Adding the male-typed symptoms, anger, substance use and risk-taking, gave men 26.3 per cent against women's 21.9. Combining the two lists gave 30.6 against 33.3, and the difference had a p of .57, which is no difference. Whether the male list measures depression or something else is contested in PS-04, where Regier is the named critic. Men die by suicide at three to four times the rate of women, a figure PS-04 gives without a primary source, and I put the two sentences side by side without knowing how far one explains the other.

Competition and risk are the other half of this section, and they are where the culture layer is most visible, because economists have run the experiment. Byrnes, Miller and Schafer pooled 150 studies of risk-taking in 1999 and found men took more risk in 14 of 16 categories, financial risk at about $d = 0.43$, social risk at −0.02, the gap shrinking over the decades of publication. Niederle and Vesterlund gave 80 people a choice in 2007 between piece-rate pay and a winner-takes-all tournament on the same arithmetic. The sexes performed equally. Seventy-three per cent of the men chose the tournament and 35 per cent of the women, with men overestimating their rank and women underestimating it.

And then the room changes and the bells move. Gneezy, Leonard and List ran the tournament choice in 2009 with the patriarchal Maasai, where 50 per cent of men and 26 per cent of women competed, and with the matrilineal Khasi, where 54 per cent of women competed and 39 per cent of men. The gap reversed with the kinship system. Saccardo, Pietrasz and Gneezy ran it on 97,678 working-age adults in 2022 and found that outside science and engineering, women's participation fell by 25.5 per cent when the task turned competitive, and that among people who had sorted into science and engineering the gap was gone. The reading offered in PS-05 is that competition-averse women have already been filtered out. The reading offered in OR-03, by way of the Khasi, is that the aversion is taught. Both can be true of the same bells. Testosterone, the hormone everyone reaches for here, gives you almost nothing: the meta-analysis of the dual-hormone hypothesis, 33 studies and 8,538 people, found the interaction between testosterone and cortisol predicts status behaviour at $r = -0.061$, a number too small to use, and PS-05 attributes it to Geniole 2016 while BI-11 attributes the same figures to Dekkers 2019, a discrepancy logged and not resolved here.

Desire is the axis where the bells sit furthest apart among the things people do together, and it is the one where the reasons are least settled. Frankenbach, Weber, Loschelder, Kilger and Friese pooled 211 studies and 621,463 people in 2022 and found men report more sexual motivation at $g = 0.69$, with a confidence interval from 0.58 to 0.81, and 0.54 after correction for the tendency of men to inflate and women to deflate. Shared area at 0.69: roughly three quarters. Conley and Yang replied in 2024 that the gap tracks what each sex gets from the act, orgasm above 90 per cent for men in partnered sex against 35 to 65 per cent for women, more pain,

and the background threat of violence, and that equalise those and the desire gap would close. Frankenbach and colleagues agreed the experiences differ and held that the trait differs regardless of why. Klusmann found in 1,865 people in 2002 that women's desire falls with the duration of the relationship and men's does not. The bells here are physiology, behaviour and culture at once, and no one has managed to pull them apart.

Attachment is the last axis and the quietest. Del Giudice pooled 113 samples and 66,132 people in 2011 and found men a little more avoidant and a little less anxious, with a multivariate distance of 0.28 in community samples, about 78 per cent shared area, and 0.12 in college samples. On the axes where people fall in love and stay, the bells are nearly one bell.

The mosaic and the classifier

Everything so far has been one bell-pair at a time. The last part of the argument is about what happens when you stop doing that, and it starts with the children who do not fit the bells at all.

Autism is diagnosed in boys more often than girls, and for decades the ratio was given as four or five to one. Loomes, Hull and Mandy pooled 54 studies and 13,784,284 people in 2017 and found that in studies that went looking, rather than waiting for referrals, the ratio was 3.25 to one, with a confidence interval from 2.93 to 3.62. Among autistic people with intellectual disability the ratio approaches two to one. Among those without it, the historical figures run from six to ten to one, and a 2024 Swedish register study found the cumulative ratio close to 1.2 to one by the age of twenty. The bell for girls is not smaller. It is diagnosed later. Kentrou and colleagues found that among 1,211 autistic adults, 31.7 per cent of women had been given at least one wrong psychiatric diagnosis first, against 16.7 per cent of men. Hirvikoski and colleagues found in 27,122 autistic Swedes in 2016 that among those without intellectual disability suicide was the leading cause of premature death, at about nine times the population rate.

Under the diagnosis sits a physiological finding I think is one of the most important in the book. Wigdor and colleagues showed in 2022, from the Danish iPSYCH cohort, that siblings of 1,707 autistic girls had higher rates of autism than siblings of 6,270 autistic boys, and Jacquemont and

colleagues showed in 2014 that autistic girls carry rare copy-number variants with a median of ten to fifteen genes where autistic boys carry two or three. It takes more genetic disruption to push a girl over the line. This is the female protective effect, and it is the bells' explanation for the male tilt of most early-onset conditions: not a male bell shifted towards disorder, but a female bell held back from it, at a cost that arrives later.

So the tail of the bell is male, on the low side and, in the greater-male-variance literature, on the high side too. Ritchie and colleagues scanned 5,216 people from the UK Biobank in 2018, 2,750 women and 2,466 men between 44 and 77, and found total brain volume larger in men at $d = -1.41$, one of the very few large bell-shifts in the book, with a variance ratio of 0.82, meaning men's brains vary more. In cognition, the variance ratio runs between 1.08 and 1.20 according to BI-11, which is enough to make both tails of the distribution male, and Harrison and colleagues found no trace of greater male variance in personality across 220 animal species in 2022. Whether the volume gap survives correction for body size is the first fight. Lise Eliot, the neuroscientist in Chicago, concluded in 2021 that once total size is removed sex explains less than 1 per cent of the brain's structural variance. DeCasien and colleagues concluded in 2022 that regional differences survive the correction at small to moderate size, particularly in the limbic system.

The second fight is the one that gives the chapter its title. Daphna Joel, the neuroscientist in Tel Aviv, took more than 1,400 scans in 2015 and, region by region, marked each as male-end, female-end or intermediate. Then she asked how many brains were all one thing. Depending on the dataset and threshold, between 1 and 6 per cent were, and BI-07 and PS-08 give the ceiling as 8. The rest were mosaics. Her conclusion was that human brains do not belong to one of two distinct categories.

Adam Chekroud and colleagues, in one of three replies published in the same journal in 2016, fed the same kind of data to a classifier and had it guess the sex of 1,166 people from brain structure alone. It was right 93 per cent of the time. Marco Del Giudice, whose multivariate distance the Preface has already explained, ran Joel's method on simulated data from animals whose faces do not overlap at all and showed the method calls those mosaics too. His point, and Chekroud's, is that fifteen bells each sharing 80 per cent of their area can, taken together, share almost none, because the traits are correlated and the classifier reads the correlations.

That is what Del Giudice, Booth and Irwing found in 10,261 Americans in 2012: warmth at $d = -0.50$, dominance at 0.27, emotional stability at 0.32, each a modest bell-pair, and all fifteen together at a distance of 2.71, 10 per cent shared area. Drop the biggest single trait and it is still 1.71, 24 per cent. Kaiser and colleagues repeated it on 31,637 people and got a comparable figure, which the Preface gives as 2.06 and BI-11 does not give at all.

I have come to think the two sides are answering different questions and both answering correctly. Joel is asking whether any one brain is a male brain. Almost none is. Chekroud is asking whether, given a brain, you can say which sex it came from. Almost always. The bells drawn one at a time overlap. The bells drawn all at once do not. The fight is not finished, and it is a fight about brains, not about kitchens.

Which brings me to the present, and to the tool. When you next hear that men are like this and women are like that, do not ask whether it is true. Ask three things. What was measured, and by asking or by instrument. What is d, and therefore how much of the two bells is shared. And is this one bell-pair or fifteen. The first question tells you which layer you are in. The second tells you what a single person in front of you is likely to be, which at $d = 0.20$ is almost nothing and at $d = 0.93$ is a seven-in-ten bet. The third tells you whether you are hearing about a trait or about a profile. Most of the claims in circulation fail the first, ignore the second, and confuse the third. The drawing on the napkin survives all three.

THE STRONGEST CASE AGAINST

The case against this chapter is that the bells are painted, not grown, and it is a strong case because the chapter's own numbers make it.

Start with the drift. Every place where the same trait was measured by asking and by instrument, the instrument found less: empathy under skin conductance and facial electromyography in Eisenberg and Lennon's 1983 review, mind-reading under a cash incentive in Klein and Hodges 2001, children's emotional display when the stranger leaves the room in Chaplin and Aldao 2013, the smile when rank is held constant in LaFrance, Hecht and Paluck 2003, and language lateralisation in Sommer and colleagues 2004, where the interval runs through zero. If the physiological layer keeps going quiet as you approach it, the reasonable inference is

that most of what the questionnaires report is culture, and that the two bells are two costumes. Alice Eagly and Wendy Wood made that argument formally in 1999, in the social role theory PS-01 and PS-05 both name: the division of labour produces the expectations, the expectations produce the self-reports, and the self-reports get mistaken for nature. The Khasi run the argument as a field experiment. Where the kinship system flips, the tournament choice flips with it.

Second, the one place the bells sit far apart, interests, has the weakest claim to a physiological origin of anything in the chapter. The people-versus-things axis is measured by inventories that ask what you would like to do, and Thomas Breda and colleagues showed in 2020 in the Proceedings of the National Academy of Sciences, with about 300,000 pupils, that the countries with the widest gaps in science choice are the countries with the strongest stereotypes about who does science. Breda's 2024 paper argues the gender-equality paradox itself is a Simpson's paradox that thins out or vanishes when countries are grouped by region. If the largest bell-shift in the chapter is a cultural product, the chapter's headline picture, most small and a few large, loses its most impressive large.

Third, desire. Conley and Yang's 2024 reply to Frankenbach is not a quibble about self-report. It is a claim that the trait being measured is the reasonable response to different experiences: partnered sex that ends in orgasm above 90 per cent of the time for one sex and 35 to 65 per cent of the time for the other, with pain and threat on one side only. A bell-shift of 0.69 that closes when the experience is equalised is not a difference in the sexes. It is a difference in what happens to them.

Fourth, the brains. Joel 2015 is the primary source and it says that between 1 and 6 per cent of brains are consistently male or female across regions. Eliot 2021 says that after size is removed, sex explains under 1 per cent of structural variance. The classifiers that rescue the categorical picture do worst, at 60 to 70 per cent, exactly when total volume is controlled, and BI-07 records the open question of whether the deep-learning models are reading organisation or cryptic size.

Put the four together: the physiological layer is nearly silent, the behavioural layer reports small differences that shrink under instrument, the large ones live in interests and desire where the cultural explanation is at least as good, and the multivariate rescue depends on brain size. A reader who holds that view is not being unreasonable. The chapter's answer, that

the differences are small but the same direction in 55 nations and present in temperament at three months, does not refute it. It only says the costume was cut to something.

IN PRACTICE

The bells are a tool for looking, not a set of instructions, and each of the following is offered as a habit to try with the failure that comes with it.

If you hear a difference claimed, convert it before you believe it. A difference reported as a percentage of one sex, or as a headline, can usually be turned into a *d* with a minute's search, and a *d* can be turned into a shared area with the table in this chapter. At 0.20, which is most of psychology, the person in front of you is a coin toss. At 0.93, which is interests, it is a seven-in-ten bet, which still loses three times in ten. The failure mode is the conversion done once and then forgotten, so that a real large difference such as interests or mental rotation gets waved away as small because the habit has become a reflex for dismissing. The tool cuts both ways, and it is supposed to.

If you are in an argument and the words have stopped working, check the pulse before you check the logic. Gottman's threshold of about 100 beats a minute is physiology, and a person past it is not stubborn, they are flooded, and the recovery takes twenty to thirty minutes whichever sex they are. The failure mode is the diagnosis used as a door: flooding is real, and "I'm flooded" said at the first hard sentence is a way of never having the conversation. The check is whether the person who calls the break also names the time they are coming back.

If you want to help someone who is frightened before something, ask which kind of help their body takes. The Trier work suggests that on average a man's cortisol falls when his partner tells him it is going to go well and a woman's does not, and that hers falls with ten minutes of hands on the shoulders, though the samples are small and the pattern is an average. The failure mode is the average applied as a rule to the person you live with, who may be the one in the other tail. The habit worth keeping is to try both, on different days, and to notice which one is followed by sleep.

If you are worried about a man's mood, look at the male-typed list as well as the standard one. Martin's 2013 figures suggest that anger, drinking

and risk-taking may carry in men what sadness and withdrawal carry in women, and that when both lists are counted the sexes come out level. The failure mode runs both ways. Reading every bad temper as depression turns a difficult man into a patient. Reading none of them as depression is the older failure, and the suicide ratio, whatever its exact figure, is the cost of it.

If you are asking people to compete for something, a promotion, a place, a raise, change the default before you change the people. Saccardo's 97,678 adults suggest that outside science the drop in women's participation when a task turns competitive is about a quarter, and OR-03's opt-out framing suggests the drop disappears when everyone is entered unless they withdraw. The failure mode is the opt-out that becomes a trap, so that people who genuinely do not want the tournament are dragged into it and punished for losing, which Buser's 2016 work suggests hits women's subsequent effort harder than men's. The check is whether withdrawing is as cheap as staying in.

If desire has gone quiet in a long relationship, treat it first as a question about the experience and only second as a question about the person. Klusmann's figures say women's desire falls with duration and men's does not, on average. Conley and Yang's figures say the average woman is getting much less out of the act. Muise's figures say satisfaction rises with frequency up to about once a week and no further. The failure mode is the schedule as contract, once a week written into the calendar as an obligation, which turns responsive desire, the kind that comes after arousal rather than before it, into a debt. The habit that survives is to change what happens, not how often.

If a daughter is quiet, capable, exhausted by school and unaccountably anxious, keep in mind that the female bell for autism and attention conditions is diagnosed later and is often misread as anxiety or a personality disorder first. Kentrou's 31.7 per cent is the figure. The failure mode is the over-reading, the quiet child pathologised for being quiet. The check is not the quietness but the cost of it: whether the competence at school is followed by collapse at home, which is what Raymaker's burnout looks like from the outside.

And if you find yourself building a policy on one bell-pair, stop and count the bells. One trait at $d = 0.20$ is nothing to organise around. Fifteen traits together may separate the sexes almost entirely, as Del Gi-

udice's distance suggests, and may separate them by nothing more than the brain's size, as Eliot suggests. The failure mode of the first is to design for a difference that does not exist in any one person. The failure mode of the second is to design as if profiles did not exist because no single trait carries them. The chapter has no third option to offer. It has only the drawing, two bells on one axis, and the discipline of asking how far apart the tips are and how much of the area is shared, before anyone decides anything.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- The conversion table from *d* to shared area and probability of superiority (BI-11, McGraw and Wong 1992).
- The drift: self-report gaps larger than observed gaps larger than instrument gaps (Chaplin and Aldao 2013, Eisenberg and Lennon 1983, Klein and Hodges 2001, LaFrance 2003).
- Interests at $d = 0.93$ as the largest replicated behavioural difference (Su 2009).
- The partner-support paradox and the cortisol asymmetry (Kirschbaum 1995, Kudielka and Kirschbaum 2005, Ditzen 2007).
- The female protective effect (Wigdor 2022, Jacquemont 2014, Loomes 2017).
- Joel 2015 against Chekroud 2016 and Del Giudice 2016 as two correct answers to two questions.

Decoration

- The Trier room in the opening.
- Mehl's word counts and the 20,000 versus 7,000 myth.
- Zimmerman and West's 96 per cent.
- The napkin drawing.

Contested

- Mental rotation *d*, 0.57 to 1.0 across reports.
- Whether Del Giudice's multivariate distance survives without latent-factor correction (Hyde's critique) and what Kaiser's replication value is.
- Whether the classifiers read organisation or size (60 to 70 per cent after volume correction versus 87 to 92 per cent volume-matched).
- Gottman's 85 per cent of stonewallers, blog-sourced.
- The gender-equality paradox (Stoet and Geary 2018 versus Breda 2020 and 2024).
- Whether Martin 2013 measures male depression or something else (Regier).

- The dual-hormone meta-analysis attribution, Geniole 2016 versus Dekkers 2019.

NUMBERS TO CARRY

- *d* = 0.20 means 92 per cent shared area and a 56 per cent chance the random draw goes the expected way. *d* = 0.80 still means 69 per cent shared. Height is *d* = 2.6 and 19 per cent shared. (BI-11 conversion table, McGraw and Wong 1992, Cohen 1988, Nelson 2012 for height.)
- Men's cortisol response to the Trier test is 1.5 to 2 times women's on average, and luteal-phase women match men. (Kudielka and Kirschbaum 2005, Kirschbaum 1999, PS-04.)
- Women 16,215 words a day, men 15,669, range 500 to 47,000, 396 students. (Mehl et al. 2007, Science, PS-03.)
- Autism male-to-female odds ratio 3.25, 95 per cent CI 2.93 to 3.62, from 54 studies and 13,784,284 people. (Loomes, Hull and Mandy 2017, PS-08.)
- Joel: 1 to 6 per cent of brains internally consistent. Chekroud: 93 per cent classified correctly from 1,166 scans. (Joel 2015 PNAS, Chekroud 2016 PNAS, BI-11 and PS-08.)

THE LIBRARY

- PS-01 Sex differences in personality, interests and values: Big Five, HEXACO, temperament, Su 2009, Cross 2011, SMPY.
- PS-02 Emotion: experience, expression and regulation by sex: Chaplin and Aldao, Kirkland, Ditzen, rumination, alexithymia, Balliet, Gottman.
- PS-03 Communication styles: Mehl, Leaper, LaFrance, Sommer, Ochs, Schrodt.
- PS-04 Stress, coping and mental health by sex: Trier test, Kudielka and Kirschbaum, Kirschbaum 1995, Ditzen 2007, Salk, Martin.
- PS-05 Competition, status, risk and hierarchy: Byrnes, Niederle and Vesterlund, Saccardo, dual-hormone meta-analysis.
- PS-06 Sexual desire and the couple: Frankenbach, Conley and Yang, Klusmann, Muise, Pastor.
- PS-07 Attachment, love and long-term bonding: Del Giudice 2011, Marazziti, Buss and DeSteno, Eastwick and Finkel.
- PS-08 Neurodiversity through the lens of sex: Loomes, Wigdor, Jacquemont, Kentrou, Hull, Raymaker, Hirvikoski, Greenberg, Ridley, Dooley, oestrogen and schizophrenia.
- BI-07 Brain differences and the mosaic debate: Ritchie, Eliot, DeCasien, Joel, Chekroud, Del Giudice, Kaiser, Stoet and Geary, Breda.

- BI-11 Within-sex variation and individuals: conversion table, Voyer 1995, Del Giudice 2012 univariate values, variance ratios, Harrison 2022, Breda 2024.
- OR-03: Gneezy, Leonard and List 2009, opt-out framing.

CHAPTER 5

Two Shelves and a Trap

STAND in the aisle of a bookshop with a paperback in each hand and let your arms hang. In the left hand, *Testosterone Rex* (Cordelia Fine, 2017), which says the idea of one molecule running the male half of the species is a myth. In the right, *T* (Carole Hooven, 2021), which says the molecule runs a great deal. They weigh about the same. Each was written by a scientist who has read the other's book and was not persuaded. Behind your left shoulder is the shelf that produced the first, with the mosaic brain, the toy aisle and the plough. Behind your right is the shelf that produced the second, with the Pleistocene, the hunters and the grandmothers. The aisle between them is about a metre wide, and this chapter would like you to stay in it for an hour.

Try it in a shop that stocks both. Notice which hand feels heavier. That is not the paper. That is the shelf you already belong to.

I belong to one. I said so in the Preface, because the confession is cheaper than pretending the man holding the books has no thumb on the scale. My mother's generation belonged to the other, with better reasons than mine, since the arguments on my shelf had been used against them in living memory. So this chapter is not going to tell you which shelf is right. It is going to show how each goes wrong, which is the same way with the sign reversed. It is going to sort what has survived a decade of large, pre-registered, unfriendly replication from what has not and from what is still being fought over. It is going to take the one fight I most want my side to win and show what a fight nobody has won looks like. And it ends with rules for reading any sentence that begins "men are" or "women

are", including the ones in this book.

The trap has two jaws. Each shelf has a favourite. Neither, in my reading, has noticed it is standing in the other one.

The gap that Hume found

In 1739 David Hume noticed that the moral writers of his day would spend pages describing how the world is and then, without announcing the change, begin saying what ought to be, as if the second followed from the first. It does not, he said, and nobody has found the bridge since. In 1903 G. E. Moore gave the mistake its name in *Principia Ethica*: the naturalistic fallacy, the step from "this is how nature made it" to "this is how it should stay". It is the hazard of the shelf behind my right shoulder. The trait has deep roots, runs the reasoning, therefore the trait has a mandate. Infidelity has deep evolutionary roots. So does killing between groups. Nobody concludes we are obliged to keep either, yet the move gets made whenever a difference between the sexes is called "natural" in a tone that expects the argument to end there.

The second jaw was named later, from an unlikely quarter. In 1978 the microbiologist Bernard Davis wrote in *Nature* about what he called the moralistic fallacy: the assumption that because a finding would be unwelcome if true, it must be false. The first jaw runs from is to ought, the second from ought to is. It is the hazard of the shelf behind my left shoulder, and harder to see, because it looks like scepticism. A finding arrives, the first question asked is not "is it true?" but "what would people do with it?", and the second question is allowed to answer the first.

Here is the twist. Each shelf is quick to spot the other's jaw and slow to spot its own, and each, looked at closely, uses both. The evolutionary shelf commits the moralistic fallacy whenever it waves away the German border, which chapter 13 will show moving a whole population of mothers by 350 hours a year, because a culture that strong is inconvenient. The constructivist shelf commits the naturalistic fallacy whenever it argues from "the brain is a mosaic" to "therefore we ought to arrange things as if it were", which is Hume's forbidden step taken backwards. Both shelves also keep a creation myth in stock, Rousseau's noble savage and Hobbes's nasty savage, and neither is what the archaeology shows.

One shelf-one story taken to the archive, which I would have told you a few years ago. Evolutionary psychology has long carried a hypothesis it calls the dual mating strategy: ancestral women secured provisioning from a reliable partner and, on the quiet, better genes from more formidable men. If so, the historical rate of children fathered by someone other than the husband should be substantial. It can be measured now. A team led by Larmuseau, working in the Low Countries, matched Y-chromosome analyses to genealogical records going back five hundred years (*Trends in Ecology & Evolution*, 2016, and *Current Biology*, 2019). Extra-pair paternity over those five centuries, well before modern contraception, ran at about 1 to 2 per cent per generation. In tightly knit farming villages under heavy surveillance it fell to about 0.4 per cent. Among the poorest classes in the crowded, anonymous cities of the Industrial Revolution it rose to about 5.9. A hypothesis that needed many cuckolds found few, and found them where the neighbours had stopped watching.

The sentence that survives

If natural does not mean right, and unwelcome does not mean false, what grammar is left? The prehistory report gives it in one line and the Preface has borrowed it: *given what we are, arrangement X under condition Y tends to produce Z.* Not "men are, therefore society must". A conditional, with the condition stated, and a verb that admits it describes a tendency in a population rather than a law for a person.

The careful authors on both shelves already write this way. Sarah Blaffer Hrdy showed that the cost of raising a large-brained, slow-growing human infant could never have been met by a mother and father alone, and that we evolved as cooperative breeders, with grandmothers and other kin doing much of the holding. She does not conclude that women ought to live in multi-generational communes. She concludes that the infant arrives expecting a network of caregivers, and that a society which isolates a mother in a flat with one child and one partner should expect the bill. Robert Sapolsky makes the same move at the level of the molecule: testosterone does not manufacture aggression, it lowers the threshold for whatever the local culture already rewards with status. Both are conditionals. Neither prescribes.

So the conditional sentence does not just avoid the two fallacies. It relocates the argument. Biology, in the report's phrase, dictates the cost of a social arrangement, not its moral worth. Say a couple wants to split the first year of a child's life exactly in half. Nothing in the endocrine accent forbids it. What the accent does is set a price: the mother has paid the whole metabolic bill for nine months and, if she nurses, is still paying it, and the father's body, as chapter 8 will show, is only now adjusting. A half-and-half year needs more scaffolding than the folklore assumes, and saying so is not the moralistic fallacy in reverse. It is a price list. Prices are not verdicts. You can pay them, and many couples do. You just should not be told they are zero.

That is the grammar for the rest of the book. Every recommendation from here on is a conditional or it is cut.

What held, what fell, what is still on the table

The Preface listed the fallen. This section is about telling in advance which way a finding is likely to go. The fallen have a family resemblance, and so do the survivors.

Take the most famous fall and look at the mechanism. In 2010 Carney, Cuddy and Yap reported in *Psychological Science* that two minutes of expansive posture raised testosterone, lowered cortisol and made people take more financial risk. In 2015 Ranehill and colleagues, same journal, ran it again, pre-registered, with 200 participants, which the report calls much larger than the original. No effect on testosterone, none on cortisol, none on risk. One thing survived: people who held the pose said they *felt* more powerful. Carney, the original's lead author, later disavowed the finding in public. The lesson is that a small unregistered study measuring a hormone, reported with a large effect that made a wonderful talk, is exactly the shape of thing that falls.

The single-sex schools fell differently. Erin Pahlke, Janet Hyde and Carlie Allison, 184 studies, 1.6 million pupils, 21 countries, 2014, allowed for who gets sent to such schools, and the advantage shrank to a tenth of a standard deviation for girls and six hundredths for boys in mathematics. A finding that depends on not asking who was in the room is the second family of the fallen.

The survivors share features too: large samples, many countries, effects that were big to begin with, and a physiology underneath. Upper-body strength, the Preface's two standard deviations, holds in every occupational cohort tested, women producing on average 60 to 75 per cent of men's. The interest gap holds: things against people, 0.93 of a standard deviation across 503,188 respondents in 47 interest inventories (Rong Su, James Rounds and Patrick Armstrong, 2009), 1.11 for engineering. The fatherhood drop holds: Lee Gettler's 394 men in the Philippines, followed for five years, and the replications since. And the argument between Hyde's 78 per cent small and Del Giudice's ten per cent overlap is not a failure of replication at all. Both sets of numbers reproduce. What differs is the question, one trait at a time or the whole profile at once, and the replication report's own verdict is that the choice of metric here remains ideological. I agree, on both shelves.

Then the middle, the largest category and the most useful. Empathy: give half a million people the Reading the Mind in the Eyes test and women do better, across ages and countries (Greenberg and colleagues, *PNAS*, 2018). Pay men to be accurate and the gap goes (Klein and Hodges, 2001). The report's grade: mixed, largely motivation.

The best example of the middle is stereotype threat, because it cuts both ways and each shelf holds one edge. In 1999 Spencer, Steele and Quinn, in the *Journal of Experimental Social Psychology*, gave able women a hard mathematics test. Told it was known to produce sex differences, they did worse. Told it was not, they matched the men. It became the constructivist shelf's flagship. In 2012 Stoet and Geary went through the replications. Fifty-five per cent of the designs reproduced the effect. Half of those had adjusted for prior mathematics scores in ways that manufacture it. Of the clean replications, 30 per cent worked. Both shelves misread that. Three clean studies in ten is a real phenomenon. It is also not what it was sold as, because a thing that happens in three rooms out of ten cannot carry a national gap in engineering. Negotiation has the same shape. Across 10,888 participants men come out of bargaining with better economic outcomes (Mazei and colleagues, *Psychological Bulletin*, 2015). Have the women negotiate for someone else, or make the range transparent, and the gap disappears or reverses. A difference that is robust and a difference that moves with the conditions are the two halves of a conditional sentence.

The paradox, worked through

I promised one fight in full. This is the one I most want my shelf to win, so it is the one I should least trust myself on.

The pattern first, in its own numbers, because the pattern is not in dispute. Costa, Terracciano and McCrae, 23,031 people in 26 cultures, 2001: women score higher on average in neuroticism and agreeableness, and the gaps are widest where sex roles are loosest. David Schmitt, Anu Realo, Martin Voracek and Jüri Allik, 17,637 people in 55 nations, 2008: the same, the gap predicted by a nation's wealth, health and education. Lee and Ashton, 347,192 people in 48 countries, 2020: an emotionality gap of 0.84 of a standard deviation, in another personality model. Falk and Hermle in *Science*, 2018, 80,000 people in 76 countries, on risk, patience, altruism and trust: gaps near zero in poor countries and largest in rich ones, correlating 0.67 with the log of income per head. And the one that made the newspapers, Stoet and Geary, 2018: 472,242 fifteen-year-olds in 67 countries, and in two thirds of them girls matched or beat boys in science. Yet Finland, Norway and Sweden, the most equal societies on the index used, produce fewer than 25 per cent women among their STEM graduates, while Algeria, Tunisia and Turkey produce 35 to 40.

My shelf's reading is elegant. Where subsistence is not at stake, people follow their interests, and interests, on average, differ. Freedom reveals. I find it persuasive, and I notice that I do.

Now the other column, which is longer than my shelf usually admits. Richardson and colleagues, 2020, in the same journal as Stoet and Geary, found that the STEM result rested on an undisclosed measure of women's "propensity" to graduate in STEM rather than on their plain share of graduates. Recompute with the share and the correlation shrinks, in some specifications to nothing. A correction was issued. Thomas Breda, with Jouini, Napp and Thebault, *PNAS*, 2020, 300,000 pupils in 64 countries: the stereotype that mathematics is not for girls is *stronger* in the richer, more equal countries, and once you allow for it the link between wealth and the STEM gap goes away. Then the study that matters most, because it asks what the cross-sections cannot. Fors Connolly, Goossen and Hjerm, 2020, took eight rounds of the European Social Survey, 235,339 people across 32 countries, and asked: as a country becomes more equal over time, does its own gap widen? The cross-section repro-

duced. The within-country answer was no. The paradox, on this evidence, is a fact about which countries are which, not about what equality does when it arrives. And Herlitz and colleagues, re-analysing 27 large studies in 2024, found the paradox holds for personality and reverses for mathematics and partner preferences. Whatever is going on, it is not one thing.

So where does that leave the worked example? The pattern is robust. The mechanism is unknown, and the replication report grades it that way. My shelf's story explains the personality data and stumbles on the stereotypes and the within-country test. The other shelf's story explains the STEM data and has to work hard on emotionality at 0.84 in 48 countries. The honest conditional, the only sentence that survives the above, is this: given that sex differences in some traits widen where constraint loosens, and given that the stereotypes about those traits widen in the same places, an institution that wants to know which it is looking at tends to have to measure both, in its own population, over time, rather than read the answer off a map of Europe. That is less than I wanted. It is what there is.

The trust rules

Which brings me to the library this book was built from, because the rules I have just used on other people's findings apply to my own, and nearly failed.

The research behind this book is machine-written: 81 reports, each from a brief that asked for the case against as well as the case for. One brief asked for the strongest statement of the shelf I do not belong to. The report that came back had read the brief backwards. It took the constructivist position as the book's own, called the biological case the "steelman" to be dismantled, and argued the constructivist case thoroughly and well, with the biological rebuttals tucked in as concessions. It is the best-argued document in the library, and it is against me.

The same library disagrees with itself. One report gives Stoet and Geary's sample as 472,242 adolescents in 67 countries, another as over 300,000 in 73. One gives Hyde's share of small differences as 78 per cent, another gives Zell's as 85.5. And the two figures this chapter most needed, the size of the power-posing replication and the residual effect of single-sex schooling, had been stripped out of the report's text by the maths

formatting, and I had to go to the page image to read them. I did not guess them. The method fits on a card.

Numbers carry names: study, author, year, sample, in the sentence. A contested finding arrives in the same sentence as the counter-study, or not at all. A fallen finding is named as fallen with the replication that felled it. Where two sources disagree, the text carries the range or a marker, never the convenient number. Where a figure is missing, it stays missing. And the case against gets its own section in every chapter, in its own numbers, even when, as here, a machine wrote it better than I would have.

Now the tool.

THE STRONGEST CASE AGAINST

The chapter has argued that the two shelves fail symmetrically and that a replication map and a conditional sentence are the way between them. The strongest case against is that the map is itself a shelf, drawn from my own side.

Start with the physiology the chapter leans on. Lise Eliot and colleagues, 2021, synthesising three decades of MRI: once you allow for the roughly 10 per cent larger male brain, sex explains about 1 per cent of the variance in structure. Rebecca Jordan-Young, 2010, went through more than 300 studies behind the theory that prenatal testosterone organises the brain and found the dose-response rule failing, with the flagship natural experiment, girls with congenital adrenal hyperplasia, confounded by years of surgery and medical stress that no control group shares. Sari van Anders, 2011, in *Psychoneuroendocrinology*, showed that testosterone behaves as a dependent variable: caring for an infant lowers it in men, competition and authority raise it in both sexes. On that reading the hormone levels of adults in a sex-stratified society are a *result* of the stratification, and a book that files them under physiology has put a cultural fact in the wrong drawer.

Then the behaviour. Janet Hyde, 2005: 78 per cent of psychological sex differences below 0.35 of a standard deviation. Zell, Krizan and Teeter, 2015, more than twelve million people: an average of 0.21, an overlap of about 85 per cent, and a random woman outscoring a random man on a "male" trait 42 times in a hundred. Mathematics across 1,286,350 people:

0.05 (Lindberg and colleagues, 2010). Verbal ability: 0.11 the other way (Hyde and Linn, 1988). The chapter's "survivors" are the exceptions this literature already lists. Alice Eagly and Wendy Wood, 2023, found that 71 per cent of the sex differences in the meta-analytic literature mirror the effects that experimental manipulations of *power* produce in people of either sex. If that holds, most of what the map calls a difference between the sexes is the psychology of the powerful and of the subordinate, and Anne Fausto-Sterling's work on the first fifteen months of life shows caregivers embedding the roles before the child can sit up. The map's categories, held, fell, contested, presuppose that what held is nature. The constructivist reply is that what held is history, replicated.

And the sharpest point. Hyde warned in 2005 that inflated claims of difference carry a social cost, because they become expectations. A chapter that offers a "price list" for equal arrangements is, on this reading, the naturalistic fallacy with better manners. It does not say the arrangement is wrong. It merely tells the couple, in advance and with a citation, that it is going to cost them, which is how the old shelf has always talked.

I have no rebuttal that closes this. I have Del Giudice's 2.71, the 14 of 16 risk domains and the 30 per cent of clean stereotype-threat replications, all of which the constructivist report itself concedes, and I have the wall not moving the fathers. That is a standoff, not a win.

IN PRACTICE

The scale here is you, alone, with a headline. None of this makes you right. It tends to make you slower, which in this subject is most of the job.

If the headline says "men are" or "women are", then ask for the effect size first, and for the overlap in the same breath. A gap of 0.2 of a standard deviation is two bells sharing 85 per cent of their area. A gap of 2 is two different instruments. The failure mode is the *d* on its own, which sounds precise and says nothing about the person across the table. The counter-evidence is Del Giudice: many small gaps taken together can separate the sexes far better than any one of them, so "all small" is not the end of the story either.

If the finding is about a hormone, then ask whether the hormone was measured or inferred. The ovulatory-shift literature counted days on a calendar and fell when someone drew blood. Power posing fell when 200

people were measured by researchers who had said in advance what they were looking for. The failure mode is the plausible mechanism, which is what a story has instead of a result. The counter-evidence is Gettler: measured, followed, replicated, and still a hormone.

If the finding is convenient for your shelf, then double whatever scrutiny you were going to give it. This is the only rule that works against you. The failure mode is the feeling of recognition, which is indistinguishable from being right. The counter-evidence is that convenient findings are sometimes true, the things-and-people gap being convenient for my shelf and still standing.

If the study compares countries, then ask whether anyone has watched one country change. The paradox holds across a map of Europe and did not appear inside any country over eight rounds of the same survey. The failure mode is reading a photograph as a film. The counter-evidence is that within-country change is slow and eight survey rounds may be too short, which my shelf says and might be true.

If the finding is about what women or girls do under a stereotype, then hold both edges. Told the test shows a sex difference, women did worse in the original study. In the clean replications it worked three times in ten. That is a real effect that cannot carry a national gap. The failure mode is choosing one of those sentences and dropping the other. The counter-evidence is negotiation, where the gap is robust and also vanishes when the woman is bargaining for someone else. “Robust” and “movable” are not opposites.

If the word “natural” appears, then ask which layer. Measurable in the body, statistical in behaviour, or variable with the culture. A sentence that does not say is moving a claim from one layer to another while you are not looking. The failure mode is answering “natural” with “so what”, which is the moralistic fallacy, or with “so there”, which is the naturalistic one. The counter-evidence is that the question is sometimes honestly hard: whether the interest gap is a body’s accent or a culture’s echo is the fight nobody has won.

If a machine read the research for you, then ask it for the case against and check whether it argued it. Mine, asked to steelman my opponents, switched sides. The failure mode is the fluent summary, which reads the same whether the figure underneath was checked, stripped or invented. The counter-evidence is that the machine also found Larmuseau, and I

would not have.

And if, after all that, the headline survives, then say it in a conditional and carry it. Given this population, this measure, this condition, this tends to happen. It is a less thrilling sentence. It is the one I could defend at my mother's table, and I have had to.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. Hume 1739 and Moore 1903 for the naturalistic fallacy, Davis 1978 for the moralistic. The conditional sentence as the book's grammar. The replication families: small, unregistered, uncontrolled findings fall, large, multi-country, physiology-backed findings hold. Ranehill 2015 against Carney, Cuddy and Yap 2010. Stoet and Geary 2012 on stereotype threat. The paradox as robust pattern with unknown mechanism, Fors Connolly 2020 as the within-country test. The library's discrepancies, carried rather than resolved by preference.

Decoration. The bookshop and the two paperbacks. The Low Countries as setting. The card the method fits on.

Contested. Whether the multivariate distance is a finding or an artefact (Del Giudice 2012 against Hyde 2005). Whether the paradox reflects freedom or stereotype (Schmitt, Falk and Hermle, Stoet and Geary against Breda, Richardson, Fors Connolly, Herlitz). Whether adult hormone differences are cause or consequence (van Anders 2011). Whether 30 per cent of clean replications is "real" or "failed", which depends on the shelf.

NUMBERS TO CARRY

- Extra-pair paternity in the Low Countries over five centuries: about 1 to 2 per cent per generation, 0.4 in surveilled farming villages, 5.9 in the poorest industrial city districts (Larmuseau and colleagues, 2016 and 2019, per SY-02).
- Power posing: no effect on testosterone, cortisol or risk in a pre-registered replication of 200 people (Ranehill and colleagues, 2015, per SY-04).
- Stereotype threat: 55 per cent of replication designs reproduced it, half of those confounded, 30 per cent of clean designs worked (Stoet and Geary, 2012, per SY-01).
- The STEM paradox: 472,242 adolescents in 67 countries, under 25 per cent women among STEM graduates in Finland, Norway and Sweden against 35 to 40 in Algeria, Tunisia and Turkey (Stoet and Geary, *Psychological Science*, 2018, corrected after Richardson 2020, per HC-07). Another report gives over 300,000 in 73 countries.

- No within-country divergence in values as equality rose: 235,339 people, 32 countries, eight survey rounds (Fors Connolly, Goossen and Hjerm, *Sex Roles*, 2020, per HC-07).

THE LIBRARY

- SY-01, *Steelman the opposite thesis*: the constructivist case in full. The case against rests on it.
- SY-02, *Nature, prehistory and the naturalistic fallacy*: Hume, Moore, Davis, the conditional sentence, Hrdy, Sapolsky, Larmuseau.
- SY-04, *Replication map: what holds and what fell*: the graded map, power posing, single-sex schooling, the metric verdict on Hyde against Del Giudice.
- HC-07, *The gender-equality paradox*: the worked example, from Costa 2001 to Herlitz 2024.

PART TWO

The Campfire Contract

CHAPTER 6

Flooding

The room is small and the chairs are the kind nobody chooses for a living room. A woman and a man sit facing each other with a low table between them, and on a finger of each hand is a small clip that reads the pulse through the skin. Under the shirt, on the chest, electrodes. On the palm, two more, reading the sweat before there is any sweat to feel. They have been asked to talk for fifteen minutes about the thing they most disagree on, and they have chosen money, which is what a great many couples choose.

At first the numbers are what you would expect of two people sitting down, somewhere in the low seventies for him, a little above that for her. Then he says a sentence that begins with "you always", and in the next room a pen on a moving strip of paper starts to draw a steeper line. Her pulse climbs through the eighties. His follows. She answers. Nothing in the room has changed, nobody has raised a voice, and the two lines keep climbing. Somewhere in the nineties something in him closes. He is still looking at her, but the look has gone flat, and the answers get shorter. When a line crosses 100 beats a minute the people in the next room make a note, because they have learnt that past this point the conversation is no longer being held by the two people having it. It is being held by their bodies.

Put two fingers on the inside of your wrist, if you like, and count for fifteen seconds. Multiply by four. At rest, in a chair, the figure is probably somewhere between 60 and 80. The number this chapter is about is 100. Nothing visible happens at 100. It is the pulse of a brisk walk, of carrying

the shopping upstairs. But when a person reaches it sitting still, across a table from someone they love, it is an alarm, and the alarm is older than the argument, older than the money, older than the two of them.

The couple in the chairs are one of 79. The year is 1980, and the two psychologists in the next room, John Gottman and Robert Levenson, do not yet know that they are going to follow these 79 couples for 14 years, or that the strip of paper is going to become the most quoted, most misquoted and most oversold instrument in the study of marriage (CO-04). This chapter is about what the strip measured, what it did not, and why the honest version is more useful than the famous one.

The number on the strip

John Gottman and Robert Levenson, psychologists then at the University of Illinois and Indiana University respectively, recruited their 79 couples in 1980 and fitted them with the sensors I have described (CO-04). The strip recorded four things: the time between heartbeats, the time the pulse wave took to travel from heart to finger, the conductance of the skin, which rises when the sweat glands open, and how much the body fidgeted. They published the first wave in 1985 in the Journal of Personality and Social Psychology and the long follow-up in 1999 in the Journal of Marriage and the Family under the title "The Timing of Divorce" (CO-04).

Gottman's word for what happens past the line was "flooding". It is a physiology claim, the first of the three layers, and it is measurable. Above the line the body has released adrenaline (epinephrine in the American papers) and cortisol, blood pressure has risen, and the parts of the frontal cortex that handle listening and weighing are being quietly switched off in favour of the parts that handle running away (CO-01, CO-04). Where exactly the line sits depends on which document you read. CO-04 gives 100 beats a minute. CO-01 gives 95, with 85 for conditioned athletes whose resting pulse is lower, and CO-01's source for that figure is a clinician's manual rather than a journal (CO-01). I carry the range, 95 to 100, and note that it was always a rule of thumb dressed as a threshold.

Here is the first sex difference, and it belongs to the second layer, the statistical one. In the 1985 sample, men flooded faster and stayed flooded longer than women (CO-04, citing Gottman and Levenson 1985). CO-01

adds that men took longer to return to baseline. From this Gottman built a behavioural account of the thing he called stonewalling, the flat look, the short answers, the retreat behind the newspaper. The stonewaller, in his reading, is not refusing to engage. His body has learnt that silence brings the pulse down, and it reaches for silence the way a hand reaches for a hot pan's handle and then lets go. CO-04 calls it escape conditioning, which is the right name.

Then comes a figure I am going to handle with tongs. Gottman has said that about 85 per cent of the stonewallers he observed were men. CO-01 cites it to the blog of a couples app, PS-02 and PS-03 to the same kind of source, and no document in my library traces it to a paper with a sample size (CO-01). So it is Gottman's clinical impression, and I report it as that. It may well be true. I cannot show you that it is.

The strip also fed the part of Gottman's work that everyone has heard of. From the coded conversations he named four behaviours that predicted trouble and called them, with a flair for marketing the field has never quite forgiven, the Four Horsemen: criticism, contempt, defensiveness and stonewalling, with contempt the most corrosive (CO-01). He also gave the world a ratio, about five positive exchanges for every negative one during conflict and something like twenty to one outside it. Both figures are cited in CO-01 to blogs rather than to a paper, so they go into the ledger as Gottman's figures, not as replicated ones (CO-01).

And then the claim that made him famous. In 1998, with Coan, Carrere and Swanson in the Journal of Marriage and the Family, Gottman reported that his model could tell which newlyweds would divorce with up to 94 per cent accuracy (CO-01). The 1999 paper gives 93 per cent (CO-03, CO-04). The two figures come from two papers, so I carry both.

Now the other column. Richard Heyman and Amy Smith Slep, psychologists at Stony Brook University, published a reply in 2001 in the same journal with a title that is itself the argument: "The hazards of predicting divorce without crossvalidation" (CO-01). The problem was not the horsemen. The problem was arithmetic. Gottman had built his model on a sample and tested it on the same sample. Any model with enough knobs can describe the data it was fitted to. Whether it can describe data it has never seen was never asked. The 1999 result rests on those 79 couples and, within them, on 21 divorces (CO-04). The extreme-group studies, where couples were picked for having the highest and lowest adjustment

scores, ran on samples of 60 to 79 couples by CO-01's count and 60 to 130 by CO-03's. When Heyman and Slep built a model of the Gottman kind and then cross-validated it, the prediction held up far less well. The figure reported for the model's out-of-sample positive predictive value is about 29 per cent. In words: of the couples the cross-validated model flagged for divorce, fewer than a third divorced.

They went further in 2002, taking a community sample rather than a laboratory one and testing the specific sex-typed sequences Gottman had described, the husband's refusal to accept influence from his wife, the wife's harsh start-up, the failure to de-escalate. Those sequences did not replicate (CO-01). The constructs did. Contempt, criticism, flooding and withdrawal remained solid risk factors for distress, in that sample as in every sample since. What fell was the choreography, the idea that the steps had a fixed sex. Heyman and Slep used the word equifinal, many roads to the same town, and that is the honest summary of the programme. Gottman found risk factors. He sold prophecy. The horsemen are real, and they ride on the Overlapping Bells (chapter 4's picture of two distributions that share most of their area), which is why no observer in a next room, with any strip of paper, can tell you what the couple in the chairs are going to do.

Who wants the change

The second thing on the strip was not a number but a shape, and it took a different psychologist to name it. Andrew Christensen, at the University of California, Los Angeles, described in 1988 the pattern he called demand-withdraw: one partner raises a problem, presses, criticises, asks for change, and the other goes quiet, changes the subject, leaves the room (CO-04). The more one pushes, the more the other retreats, and the retreat is the thing that makes the next push harder. You can see the pulse inside it. The withdrawer is often the one past the line.

How big is it, and who stands where? Paul Schrodt, Paul Witt and Jenna Shimkowski put the literature together in 2014 in Communication Monographs: 74 studies, 14,255 participants (CO-04). The pattern's correlation with relationship dissatisfaction was r = .360 across all the studies (CO-04). Split by who does what, the wife-demand, husband-withdraw

form was the more common and correlated with dissatisfaction at r = .380. The husband-demand, wife-withdraw form was a substantial minority and correlated at r = .392, which is to say, slightly more (CO-04). In clinical samples the figure rose to .413. Three things follow. The pattern hurts. The pattern is more often the woman asking and the man leaving. And when it runs the other way it hurts at least as much, which rules out the comfortable story that it hurts because women ask.

So why is it more often her asking? Here Christensen did the experiment that moves the question from the second layer to the third. In 2006, in the Journal of Consulting and Clinical Psychology, with Eldridge, Catta-Preta, Lim and Santagata, he and his colleagues compared couples in Brazil, Italy, Taiwan and the United States and found the same asymmetry in all four (CO-04). That looks like a body story. But CO-04 also reports, under the same reference, a study of 63 heterosexual, gay and lesbian couples in which the demander turned out to be whoever wanted more change in the relationship, regardless of sex (CO-04). I cannot tell from my documents whether these are one paper or two. What I can say is the finding: strip the sex out of the pairing and the pattern survives, attached now to the person with the complaint. The partner who wants the house to be different pushes. The partner content with the house as it is withdraws, because the partner who wants nothing changed wins by saying nothing. In the heterosexual majority it is more often the woman who wants something changed, and chapter 7 says what. That is a culture claim, layered on a physiology claim, the pulse, and a behaviour claim, the retreat. CO-04's own summary puts it in a line I am happy to borrow: biology provides the mechanism, flooding, and culture decides who demands and who withdraws.

Let me say plainly what this book assumes. The heterosexual couple is its central case because that is where the endocrine difference and the cultural script sit in the same bed. Same-sex couples are the control group. Where they show the same pattern, the pattern is not about the sexes. Where they do not, it might be.

The control group's own strip comes from the same laboratory. Gottman, Levenson and colleagues published in 2003, in the Journal of Homosexuality, a twelve-year study of 21 gay and 21 lesbian couples, coded and measured in the same way as the heterosexual sample (CO-06). CO-04 gives this study 106 couples. That is wrong: 106 is the size of a later,

uncontrolled therapy study by Garanzini and colleagues in 2017, which CO-06 reports separately (CO-06). I carry 42. In the 42, same-sex couples opened disagreements with more affection and humour, showed less belligerence, domineering and fear, and the unhappy ones showed abnormally low arousal in conflict rather than the high arousal of unhappy heterosexual couples (CO-06). Forty-two is a small number, and so says Michael Rosenfeld of Stanford University, who runs the How Couples Meet and Stay Together survey, a nationally representative panel whose 2017 wave had 3,510 respondents (CO-06). Rosenfeld's 2014 analysis found that once you control for whether a couple is married, something same-sex couples were for most of the period not permitted to be, the break-up rates of same-sex and different-sex couples are nearly identical (CO-06). The control group, in short, floods less and splits at the same rate. The pulse is not the whole of it.

The sex that reacts more depends on the measure

Everything so far has measured the heart. Janice Kiecolt-Glaser, a psychologist at the Ohio State University College of Medicine, measured the glands, and the sex difference turned around.

Her 1996 paper in the Journal of Consulting and Clinical Psychology asked in its title whether men were really more physiologically affected by marital conflict than women (CO-04). With Tamara Newton she reviewed the evidence in 2001, including a study of 90 newlywed couples, and the answer on the hormones was no. During and after conflict, women showed larger and more persistent rises in cortisol, adrenaline and noradrenaline (norepinephrine in the American papers) than men did, and the rises were largest when the husband withdrew (CO-04). Put the two laboratories side by side and you get a sentence to carry through the book. On the heart, men react more. On the adrenal glands, women do. The sex that reacts more depends on the measure, and on who has just left the room.

Then she did something that nobody who reads the paper forgets. In 2005, in the Archives of General Psychiatry, with 42 married couples, her team raised small blisters on each person's forearm by suction, and then had the couple argue for 30 minutes. The blisters of the couples who ar-

gued with hostility healed at 60 per cent of the rate of the others, and the hostile couples' blood carried more interleukin-6 and tumour necrosis factor alpha, the signalling proteins of inflammation (CO-04). A bad argument is a small wound that keeps other wounds open. That is a first-layer claim, measured.

The hormone of the chapter is cortisol, and chapter 3 has already introduced it, so I only remind you of its Endocrine Accent (chapter 3's phrase for a hormonal tilt that colours a response without choosing it). What the couple literature adds is that the accent is spoken differently by the two sexes and heard differently by their partners. Beate Ditzen, then at the University of Zurich, gave 47 heterosexual couples, 94 people, a nasal spray of either oxytocin or a placebo before a standard conflict discussion, and reported in Biological Psychiatry in 2009 that the oxytocin couples produced a higher ratio of positive to negative behaviour ($F = 4.18$, $p = .047$) and lower salivary cortisol afterwards ($F = 7.14$, $p = .011$) (CO-04 PDF, page 3). A p of .047 lives next door to chance, and the Preface has already told you how the "cuddle hormone" story fared when it was replicated. Her 2013 paper in Psychoneuroendocrinology matters more here: the same spray lowered salivary alpha-amylase, a marker of the sympathetic nervous system, in women and raised it in men, and in men the rise went with more positive communication (CO-03). One molecule, two sexes, two directions, and in neither a verdict.

How much does any of it matter over a life? Theodore Robles and colleagues answered in 2014 in Psychological Bulletin with 126 articles and more than 72,000 people: marital quality predicted health at $r = .07$ to $.21$ (PS-07), with mortality risk at $r = .11$ and cardiovascular reactivity at $r = -.13$ in CO-04's account (CO-04 PDF, page 3). Those are small numbers, comparable, the papers note, to the effects of diet and exercise (CO-01, CO-04). The German KORA-Age cohort adds the most domestic finding in the file: married men who described themselves as lonely had a flatter daily cortisol curve, the signature of a stress system that has stopped resetting at night (CO-01). A marriage is not a shelter by being one. It is a shelter when the person in it does not feel alone.

The ledger

This section is about a pressure that does not come from the partner at all. It comes from the payslip, and it is the cleanest demonstration in this book of a third-layer effect with a first-layer readout.

Lamar Pierce, Michael Dahl and Jimmi Nielsen published in 2013, in Personality and Social Psychology Bulletin, a paper called "In Sickness and in Wealth" (CO-07). They had the pharmacy records of more than 200,000 married Danish couples from 1997 to 2006, which in Denmark you can have. They lined the couples up by the share of the household income the wife earned and looked at what happened to prescriptions right at the point where her income passed his. Where it did, prescriptions for erectile dysfunction medication rose by 0.06 percentage points, a tiny number at a sharp edge ($p < .01$), and the wives at the same edge collected more insomnia and anxiety prescriptions (CO-07). The design compares couples just either side of the line, who differ in nothing but which side they are on. Two features tell you which layer it lives in. There was no discontinuity among cohabiting couples. And there was none among couples in which the wife had already out-earned the husband before they married (CO-07). The pill is physiology. The line is a contract, and it only bites the men who signed the old one.

Joanna Syrda found the same shape in American data in 2020, in the same journal, in 6,035 households of the Panel Study of Income Dynamics: men's distress traced a U against the wife's share of income, lowest at about 40 per cent and highest at 100 (CO-07). Not lowest at zero.

The edge shows up in behaviour too, and here the two-sided reading matters. Marianne Bertrand, Emir Kamenica and Jessica Pan reported in 2015 in the Quarterly Journal of Economics that the distribution of couples falls off a cliff just past the point where the wife earns half (CO-07). Then Ariel Binder and John Lam showed in 2022 in the Journal of Human Resources that the cliff vanishes if you remove the roughly 0.25 per cent of couples who earn exactly equal amounts (CO-07). The cliff was an artefact. But the lie was not. Marta Murray-Close and Misty Heggeness matched what people told the Census to what they told the tax office: when a wife just out-earns her husband the couple is 15.9 percentage points more likely to misreport, the man rounding himself up by 5.8 and the woman down by 5.1 (CO-07). Nobody does that for a line that

does not exist for them.

And infidelity. Christin Munsch used the National Longitudinal Survey of Youth, 2,750 people, in the American Sociological Review in 2015: men entirely dependent on a wife's income had an annual infidelity rate of about 15 per cent against about 5 per cent for entirely dependent women, with men's rate lowest at about 70 per cent of household income (CO-07). "On average" did a lot of work in that paragraph, so let me pay for it now. Eighty-five per cent of those entirely dependent men did not cheat in a given year (CO-07). The curve is real and most men on it are faithful.

Does any of this end marriages? Alexandra Killewald, in the American Sociological Review in 2016, took 46 years of the Panel Study of Income Dynamics and more than 6,300 couples and asked what predicted divorce in two eras (CO-06). For couples married before 1975, a household in which housework was split half and half was 36 per cent more likely to divorce than one in which the wife did three quarters of it (CO-06). For couples married after 1975, housework lost its power to predict anything (CO-06). What replaced it, for the later marriages, was whether the husband had a full-time job: without one, the annual divorce risk was 3.3 per cent against 2.5 per cent, which is a third higher, 33 per cent in CO-06 and CO-07 and about 30 per cent in CO-02 (CO-06, CO-07). The wife's income made no difference either way. I was born in 1975, the year Killewald's line falls, and my parents' household was the kind in which the old contract still held. The script changed around the children of that year while the men it bound still worked under it.

Is the script universal? Pilar Gonalons-Pons and Markus Gangl found across 29 countries in 2021 that a husband's unemployment raised the risk of separation only where the breadwinner norm was strong (CO-07). Within one country, Maximilian Sprengholz and colleagues found that a wife out-earning her husband cost life satisfaction in West Germany and nothing in the East, where she had been expected to work for forty years (CO-07). The pills are first layer. Everything else in this section is third. Nothing in between is a sex difference. It is a difference in who was promised what.

The tide and the bond

Two more things end couples, and the documents are a good deal less sure about them than the headlines are.

The first is desire, and the cleanest number comes from Hamburg. Dietrich Klusmann published in 2002, in the Archives of Sexual Behavior, data from 1,865 people in which women's reported desire fell with the length of the relationship and men's did not (PS-06). It is a cross-sectional study, so it compares different women in different relationships rather than following anyone, and PS-06 says as much. The larger claim behind it, that men on average want sex more, was tested in 2022 by Julius Frankenbach and colleagues in Psychological Bulletin across 211 studies and 621,463 people: g = 0.69, 95 per cent confidence interval 0.58 to 0.81 (PS-06 PDF, page 5). A difference of that size leaves the two distributions sharing about 73 per cent of their area. Adjusted for the tendency of men to overstate and women to understate, the figure fell to g = 0.54, about 79 per cent overlap, and did not vanish (PS-06 PDF, page 5). Then the other column, which the authors themselves printed. Terri Conley and Verena Yang replied in 2024 that women on average get less out of sex: men in heterosexual encounters reach orgasm at rates above 90 per cent and women at 35 to 65 per cent, and women carry more pain and more fear (PS-06). Frankenbach and colleagues conceded that this almost certainly depresses women's reported desire, and held that the measured difference exists whatever its origin (PS-06). Both are true. They are not a contradiction. They are the subject. For the couple in the chairs the practical findings are Amy Muise's: once a week is roughly where the well-being gain from frequency levels off (Muise, Schimmack and Impett 2016), and the total desire in a couple matters more than how well matched it is (Muise et al. 2020) (PS-06).

The second is attachment, which is where the couple literature keeps its one well-measured sex difference in temperament. Marco Del Giudice pooled 113 samples and 66,132 people in Psychological Bulletin in 2011: men scored higher on avoidance and lower on anxiety, with a multivariate difference of D = .28 in community samples, about 78 per cent overlap, .12 in student samples and −.07, meaning nothing, in web samples (PS-07). The difference is there and it is small and it depends on who answered. It also moves with age. William Chopik, Robin Edelstein and Kevin Grimm

tracked 628 people across 59 years and found anxiety peaking in young adulthood and avoidance falling in a straight line thereafter (PS-07). The avoidant man of thirty is, on average, a less avoidant man of sixty. Sometimes the same marriage gets to find out.

What, in the end, holds? Samantha Joel and a large team of co-authors put 43 datasets and more than 11,000 couples through machine learning in 2020 in the Proceedings of the National Academy of Sciences, and the result is the most useful number in Part II (CO-01). Variables about the relationship itself explained about 45 per cent of how good people said it was now. Variables about the individuals, their personalities, their attachment styles, their sex, explained about 21 per cent. And for how the relationship changed over time, nothing in either set explained more than 15 to 20 per cent (CO-01). The single strongest predictor was whether a person believed their partner understood and responded to them, which PS-07 calls perceived partner responsiveness. Not sex. Not testosterone. Not the pulse. Whether, when the pulse climbs, there is someone across the table who notices.

Which brings me back to the chairs. Everything here was measured in a laboratory, and you do not have one. But you have a wrist. The pulse is the one instrument in the whole couple literature that a person can read alone, in the middle of the argument, without anyone's permission. Count it. That is the tool.

THE STRONGEST CASE AGAINST

The case against the chapter is that I have told a body story about a contract. Here it is, made as strongly as I can.

Start with Christensen's own experiment. Put 63 couples of three kinds in the same room and the demander is whoever wants change, and sex drops out of the equation (CO-04). Add Heyman and Slep's 2002 replication, in which the sex-typed sequences, the husband refusing influence and the wife starting harshly, failed to appear in a community sample while the sex-neutral constructs held (CO-01). On this reading flooding is a human response to being asked to change, the person asked is more often the man because the person asking is more often the woman, and she is more often the one asking because she is more often the one with

the complaint, which chapter 7 explains. Nothing endocrine is needed beyond the fact that people have adrenal glands.

Then take the ledger. Pierce's Danish pharmacies show the clearest first-layer readout in the chapter, and it vanished in cohabiting couples and in couples where she had always earned more (CO-07). Gonalons-Pons and Gangl's 29 countries show the breadwinner effect switching on and off with the local norm, and Sprengholz shows it switching off at the old inner German border (CO-07). If the men's bodies were responding to being out-earned as such, the Danish cohabiting men would have been at the pharmacy too. They were not. They were responding to a promise, and a promise is a sentence, not a hormone.

Then the control group. Rosenfeld's representative data say that same-sex and different-sex couples, once you allow for marriage, dissolve at the same rate (CO-06). If the pulse were destiny, the couples that flood less should last longer. They do not, measurably.

And then Joel. Twenty-one per cent of present quality for everything about the individuals combined, sex included, and no more than 15 to 20 per cent of change over time for anything at all (CO-01). On this reading the Overlapping Bells are not a qualification to the chapter. They are the chapter, and the sex difference is a thin stripe on a map that is mostly water.

I think the case is largely right, and I have written the chapter so that it can be. What it does not do is make the pulse go away. Kiecolt-Glaser's blisters healed at 60 per cent of the rate whoever had signed what (CO-04). Flooding is the mechanism whichever culture is operating it, and a mechanism you can count on your wrist is worth more to the couple in the chairs than a norm they cannot see. Both are true. The contract decides who floods. The flood decides what happens next.

IN PRACTICE

Everything here is conditional, because everything above was measured on averages, and you are not one. If you can learn to notice your own pulse in an argument, you have the one tool from the laboratory that works without it. Two fingers on the wrist, fifteen seconds, multiply by four. If the figure is above 95 to 100 and you are sitting down, the conversation has moved from your mouth to your glands, and nothing said

in the next ten minutes is likely to be heard (CO-01, CO-04). The failure mode is that counting is itself a withdrawal, and a partner who sees you looking at your watch mid-sentence may reasonably take it as the newspaper by another name. So say what you are doing, before you need to.

If you are the one who goes quiet, the documents' time-out is 20 to 30 minutes, long enough for adrenaline and cortisol to clear and the pulse to come down (CO-01, CO-04). Both sources for that figure are clinicians' blogs, so treat it as a rule of thumb. The failure mode is the time-out that never ends. A pause that is not followed by a return at a stated time is stonewalling with a clock on it, and the partner left in the room, whose adrenal glands Kiecolt-Glaser tells us are now doing the reacting, is being flooded by the silence (CO-04). The pause has to carry a time, and the time has to be kept.

If you are the one who pushes, consider that the demand-withdraw pattern is held in place by both ends, and that it correlates with dissatisfaction at r = .360 whichever of you is doing which (CO-04). The question underneath it, in Christensen's reading, is who wants change, and it can be put on the table as a question rather than a complaint. "I want this to be different and you do not" is a sentence a man with a pulse of 90 can still hear. The failure mode is that naming the asymmetry does not resolve it. The person who wants nothing changed still wins by waiting, and no amount of good phrasing moves a partner who has decided the house is fine.

If you keep score, Gottman's five to one is a serviceable shape for a week and a terrible one for an evening (CO-01). The failure mode is the ledger itself. A partner who hears "that is the third negative today" is being audited, and audits are contempt in a tidy shirt.

If you are a man whose wife earns more than you, and it bothers you, the Danish data say the bother is real, measurable and most likely to be found in men who married under the older contract (CO-07). The practice is to say which contract you thought you were signing, since the discontinuity vanished in couples who had already settled the question before the wedding. The failure mode is the opposite case: if it does not bother you, there is nothing here to fix, and importing the problem because a chapter said men have it is a way of acquiring a norm you were lucky enough to be born without.

If desire has fallen on one side, Muise's finding that the total matters

more than the match suggests the first question is not "why do you want less" but "what would make either of us want more" (PS-06). Conley and Yang's reply to Frankenbach is a practical document in this light: if sex is reliably less rewarding for one partner, which the orgasm figures say it is for women on average, then the fall in her desire is information about the sex, not about her (PS-06). The failure mode is scheduling as duty. Responsive desire needs something to respond to, and a diary entry is not it.

If one of you is the avoidant one, Del Giudice's D = .28 says it is slightly more likely to be the man and Chopik's 59 years say it is likely to ease (PS-07). The practice is to read avoidance as a weather report rather than a verdict. The failure mode is patience as a substitute for change: a line falling over decades is a long time to wait, and the KORA-Age men were married and lonely at once (CO-01).

And if none of this fits, that is the Overlapping Bells at work. The husband who pushes and the wife who leaves are a substantial minority at r = .392, which is not a footnote (CO-04). Everything in this chapter is a risk factor, and Heyman and Slep's 29 per cent is the number to remember when a book, including this one, tells you what your marriage is going to do.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- Flooding as a measured state above roughly 95 to 100 beats a minute, with adrenaline and cortisol release and reduced frontal control (CO-01, CO-04). Physiology layer.
- Men flooded faster and longer in Gottman and Levenson's 1985 sample. Women showed larger, more persistent rises in cortisol, adrenaline and noradrenaline in Kiecolt-Glaser's, especially when husbands withdrew. The sex that reacts more depends on the measure (CO-04).
- Demand-withdraw: 74 studies, 14,255 participants, r = .360 with dissatisfaction, wife-demand majority at .380, husband-demand minority at .392 (CO-04). The demander is whoever wants more change (CO-04).
- Gottman's predictors are risk factors, not prophecies: same-sample validation, 79 couples, 21 divorces, sex-typed sequences failed to replicate (CO-01, CO-03, CO-04).
- Hostile conflict slows wound healing to 60 per cent of the rate in 42 couples (CO-04).

- Pierce 2013: the income discontinuity bites married men and not cohabiting men or men whose wives always earned more (CO-07). Culture layer with a physiology readout.
- Joel 2020: relationship variables 45 per cent, individual 21 per cent, change over time at most 15 to 20 per cent (CO-01).

Decoration

- The 20 to 30 minute time-out (blog-sourced in CO-01 and CO-04).
- Five to one and twenty to one (blog-sourced in CO-01).
- Syrda's U-shape and its 40 per cent minimum (CO-07).
- Munsch's infidelity curve and the 85 per cent who do not (CO-07).
- Klusmann's cross-sectional desire decline (PS-06).

Contested

- The flooding threshold: 95 bpm (CO-01) versus 100 bpm (CO-04).
- 85 per cent of stonewallers male: Gottman's clinical claim, traced in every document only to a couples-app blog (CO-01, PS-02, PS-03).
- 93 per cent (1999) versus 94 per cent (1998) prediction accuracy (CO-01, CO-03, CO-04), and the about 29 per cent cross-validated figure from SY-05, not in the CO documents.
- Gottman and Levenson 2003 sample: 42 (CO-06) versus 106 (CO-04, which is Garanzini 2017's therapy sample).
- Whether Christensen 2006's four-country study and the 63-couple heterosexual, gay and lesbian comparison are one paper or two (CO-04).
- Killewald's husband-unemployment hazard: 33 per cent (CO-06, CO-07) versus about 30 per cent (CO-02).
- Bertrand's cliff (CO-07) versus Binder and Lam's dissolution of it (CO-07).
- Frankenbach's g = 0.69 and Conley and Yang's reading of it (PS-06).
- Ditzen 2009's oxytocin effects at p = .047 against the replication record (CO-04, Preface).

NUMBERS TO CARRY

- 79 couples, 21 divorces: the sample behind the 93 per cent (CO-04). About 29 per cent positive predictive value when cross-validated, from SY-05.
- 95 to 100 beats a minute: the flooding threshold, carried as a range (CO-01, CO-04).
- r = .360, .380, .392: demand-withdraw and dissatisfaction overall, wife-demand, husband-demand, from 74 studies and 14,255 participants (CO-04).

- 0.06 percentage points: the rise in erectile dysfunction prescriptions at the point where a Danish wife's income passed her husband's, in more than 200,000 couples, absent in cohabiting couples (CO-07).
- g = 0.69, confidence interval 0.58 to 0.81, about 73 per cent overlap: the sex difference in sex drive across 211 studies and 621,463 people, falling to g = 0.54 adjusted for response bias (PS-06 PDF, page 5).

THE LIBRARY

- CO-01: What makes heterosexual couples last. Gottman, Heyman and Slep, Joel 2020, Robles 2014, KORA-Age.
- CO-03: The polarity literature. Ditzen 2013, the 60 to 130 range.
- CO-04: Conflict. The 79 couples, flooding, Schrodt 2014, Kiecolt-Glaser, Ditzen 2009, Christensen.
- CO-06: Control groups. Gottman and Levenson 2003, Garanzini 2017, Rosenfeld, Killewald 2016.
- CO-07: Money and status. Pierce 2013, Syrda 2020, Bertrand 2015, Binder and Lam 2022, Murray-Close and Heggeness 2018, Munsch 2015, Gonalons-Pons and Gangl 2021, Sprengholz 2022.
- PS-06: Desire. Klusmann 2002, Frankenbach 2022, Conley and Yang 2024, Muise 2016 and 2020.
- PS-07: Attachment. Del Giudice 2011, Chopik 2019, Joel 2020, Robles 2014.

CHAPTER 7

The Second Shift That Wasn't

It is twenty past six in the evening and a woman is standing at a cooker in a flat somewhere in Germany with a wooden spoon in one hand and a pen in the other. The pen is the unusual part. On the table behind her is a booklet from the Federal Statistical Office that cuts her day into slices of ten minutes, and every slice has to be filled in: what she was doing, what else she was doing at the same time, who was with her. Stirring the sauce goes in the first column. Listening for the child in the bath goes in the second. The pen hovers over the slice from six to ten past, because in that slice she also found the swimming kit, answered a message from the nursery and remembered that the washing machine finished an hour ago. There is no column for remembering. In 2022, 20,000 people in Germany filled in a booklet like that, and what they wrote is the closest thing the country has to a ledger of how a household spends its minutes.

Try it yourself for one day. Not the official version, just a sheet of paper with the hours down the side and a rule: write what you were doing, and what else. Most people who try it learn two things by bedtime. The first is that they do more than they thought. The second is that their partner does too, and that the two totals are closer than either had been prepared to admit at the previous argument. The third thing, which takes longer to see, is that the two columns do not hold the same kinds of minutes, and that a ledger which balances at the bottom can still be unfair on every line.

This chapter is about that ledger: who keeps it, what it says, where it lies, and what the body adds to it that the pen cannot catch. It is where I have to tell you, in some embarrassment for my own sex and some for

the other, what the diaries actually say about the famous second shift. It is not what either shelf expected.

The month that was not there

In 1989 the sociologist Arlie Hochschild published *The Second Shift*, a book built on interviews with and observation of 50 dual-earner couples, and gave the late twentieth century one of its most durable sentences. Women had gone out to work, she said, and come home to a second shift of housework and childcare, and the arithmetic of that double day added up, across a year, to an extra month of twenty-four-hour days. The revolution had stalled at the front door. For a generation the sentence was simply true. My generation grew up inside it, and it had the great advantage of all good explanations: nobody had to count anything to believe it.

Then people counted. The instrument is the time-use diary, and it is worth a paragraph because the whole chapter stands on it. Ask people how many hours a week they spend on housework and they estimate, and when they estimate they flatter themselves. A diary does not ask for an estimate. It asks what you were doing at ten past six, and then at twenty past, through the whole day, usually in slices of ten minutes, and what else, and who was in the room. The Multinational Time Use Study has been collecting diaries like that since 1961 and now holds more than 1.2 million diary days from 30 countries. It is the statistical layer, and for once the statistics are good.

In 2019 Jonathan Gershuny and Oriel Sullivan, who work from that archive, summarised fifty years of it across 14 rich countries in the *Annual Review of Sociology*. The headline is the one Hochschild's sentence cannot survive. Add paid work to unpaid work and men and women do the same amount: 450 to 500 minutes a day, about eight hours, for both sexes, and that total has barely moved in four decades. The extra month is not in the diaries. In the big samples the leisure gap is, on Gershuny and Sullivan's reading, an artefact of early measurement or of particular sub-samples, and Hochschild's fifty couples were a particular sub-sample.

What the diaries do show is a different thing and, I think, a more interesting one. Inside the same eight hours, women do about 60 per cent of the unpaid work and 40 per cent of the paid, and men the reverse. Paid

work has converged across countries to a band of 250 to 300 minutes a day. Unpaid work runs at 180 to 250 minutes and is converging slowly: women do less of it than their mothers did, men do slightly more than their fathers. And still, in every country in the archive, the routine tasks sit mostly with her. The total is equal. The kind is not.

Germany keeps its own ledger. The Federal Statistical Office's time-use survey of 2022, the booklet from the opening, with 20,000 people, puts the country's Gender Care Gap at 43.4 per cent: women over 18 do on average 1 hour 16 minutes more unpaid work a day than men, nearly 30 hours a week against 21. In 2012 and 2013 the gap had been 52.4 per cent, or 1 hour 27 minutes. So it is closing, slowly, and the routine part closes slowest. Then the survey splits the sample by children. Mothers of children under six do about 10.5 hours a week less paid work than women without children. Fathers of the same children keep their full-time jobs and trim their leisure a little. Which is how an equal ledger produces an unequal life, and how the "double burden" of my mother's generation in the West, the part-time job plus all of the housework, turns out to have been an arrangement rather than a law of nature.

The kinds of minutes

So why, if the totals match, do working mothers in every rich country report more exhaustion, more time famine and more conflict between work and home than working fathers do? There are two answers in the research, and both are about the texture of a minute rather than its number.

The first is density. In 2011 Shira Offer and Barbara Schneider took a different instrument to 402 mothers and 291 fathers in dual-earner families in the 500 Family Study: a pager that went off at random moments and asked what they were doing, what else, and how they felt about it (*American Sociological Review*). Mothers were doing two things at once for 48 hours a week. Fathers for 39. Nine hours is not nothing, but the number is the smaller finding. The larger one is what the two things were. When a mother multitasked it was childcare on top of housework, and it came with distress. When a father multitasked it was more often leisure or paid work on top of something, and it did not. The same slice of the diary, "cooking", is a different slice if someone is in the bath upstairs. Fragmen-

ted time restores less than whole time, and a diary that counts minutes cannot see the fragmentation, which is why the ledger balances and the woman holding it does not feel balanced.

The second is thought. In 2019 Allison Daminger published, in the same journal, 70 interviews with the members of 35 couples and gave the invisible part of housework a structure. Every household task, she found, has four stages. Someone has to anticipate the need before it is urgent: the shoes the child will outgrow by autumn. Someone has to identify the options. Someone has to decide. And someone has to monitor whether the thing was actually done. Deciding, the stage that carries authority and is visible to both partners and to the in-laws, was shared almost equally. Anticipating and monitoring, the stages that happen inside one head, continuously, with no slice in any diary, were done by the woman, even in couples who split the physical chores down the middle. He decides which summer camp. She noticed there was going to be a summer, and she checks that the form went in. Daminger called it cognitive labour. The internet calls it the mental load, and for once the internet is right.

Whether it hurts is a separate question with a first answer. In 2024, 322 mothers of young children were surveyed on who did the thinking in their household (Aviv and colleagues, *Archives of Women's Mental Health*). The split of cognitive labour was more skewed than the split of physical tasks, and the skew tracked maternal depression, burnout and worse relationship functioning. It is one survey, and a correlation between carrying the thinking and feeling the weight of it could run either way. It is the statistical layer, lightly populated. But it points where the pager pointed.

Now the German family panel, which the Preface quoted. Its figures on who is primarily responsible for laundry, cleaning and cooking are the ones you have seen. The figure you have not seen is the other column: men are primarily responsible for outdoor tasks and repairs in 80 per cent of couples. Put the two columns side by side and the difference is not in volume, it is in rhythm. Laundry recurs every two days and cannot wait for a convenient Saturday. A dripping tap can, and usually does. The tasks that went to women are the ones that interrupt, and if you remember chapter 2 you will recognise the oldest clause of the Campfire Contract: interruptibility, not strength, decided who did what around the fire. The modern couple has kept the clause and changed the fire.

The panel also shows when the clause arrives. Couples who begin sym-

metrical, both working, both cooking, almost invariably revert to an unequal split at the birth of the first child, and they do it on both sides of the old border. The mechanism is in the German diary above: she drops ten hours of paid work, he drops an hour of football, and the routine tasks fall into the ten hours she now has and do not climb back out when the paid hours return. Of the women in the resulting "female-overburdened" households, about one in five says she is thinking about leaving. And then the finding I keep turning over: in fixed-effects models, which compare each man with himself over time rather than with other men, a man's life satisfaction goes up when he spends more hours on routine domestic work and fewer on paid work. He likes the laundry more than the breadwinning, it appears, once he does enough of it to find out.

Which raises the obvious question, and the chapter has to answer it before it goes near the body. If the totals are equal, the kinds are what hurt, the reversion is cultural enough to be absent in half of one country, and men are happier doing the routine work, why does the ledger not simply rebalance itself?

What the ledger costs

Because there is a second ledger behind the first, and it is kept in money and in marriages.

Start with the marriages. Alexandra Killewald followed more than 6,300 American couples through the Panel Study of Income Dynamics from 1968 to 2013 and asked whether the division of labour predicted divorce (*American Sociological Review*, 2016). For couples married before 1975, a wife in full-time work raised the risk, as the old economics of specialisation said it would. For couples married after 1975, the year I was born, it did not: a wife's job and her own income no longer destabilised anything. What predicted divorce in the modern cohort was the husband. A husband not in full-time employment carried a divorce hazard roughly 30 per cent higher, or, in the other report of the same paper, 3.3 per cent a year against 2.5, a third higher. The requirement that she stay home has gone. The requirement that he earn has not. Which means that the man in the German panel who cuts his paid hours and feels better is, in the American panel, raising his odds of being left, and the two findings do

not contradict each other. They are the two halves of the ledger. The West German half, from the Preface, is the rise in a wife's housework in the year after her income passes her husband's, which does not happen in the East. Same bodies. Different ledger.

Then the money, which compounds. In Germany nearly 45 per cent of employed women work part-time against about 11 per cent of men, and because the pension follows lifetime earnings, older women draw on average 53 per cent of the pension income of older men. The ten hours a week that a mother drops at the first birth are, forty years on, half a pension, and the welfare state my parents' generation built to protect the family from the market does the compounding.

And the bed, which the sociologists could not leave alone. In 2013 Sabino Kornrich, Julie Brines and Katrina Leupp reported in the *American Sociological Review*, from about 4,500 couples surveyed between 1992 and 1994, that couples with a traditional split of chores had sex more often than egalitarian ones, and that husbands who did more of the cooking, cleaning and shopping reported less of it, about 5.2 times a month. Three years later Daniel Carlson, Sarah Hanson and Michelle Frisco took a survey from 2006, 932 couples, and found the sign had flipped: couples who shared the routine housework equally had sex 6.8 times a month, half a time more than the traditional couples and twice as often as couples where the man did most of it (Journal of Marriage and Family, 2016). Two surveys, fourteen years apart, opposite answers, and the honest reading is that the script changed between them. Then the German family panel, 1,315 couples followed over time, took the question apart (Johnson, Galambos and Anderson, 2016). Hours did not predict sex. A fifty-fifty split did not predict sex. What predicted it, for both partners, was whether she judged the split fair and whether he judged his own contribution fair. The ledger matters to desire. It matters through the sense of justice, not through the count.

Same-sex couples are the control group here, and they behave like couples who were never handed the script. They divide the work more equally than different-sex couples do, though not fifty-fifty: by time, by resources, by who minds which job less (Goldberg 2013, Umberson 2015). The part worth stealing is not the split. It is that they talk about it, because nothing tells them in advance who does the laundry, and they report their arrangements as fairer even when the hours are not equal. The

fairness paradox, as Gerold Mikula called it in 1998, is that objective inequality predicts dissatisfaction poorly and perceived unfairness predicts it well. Good, because a couple can reach fairness without reaching equality. Bad, because a couple can also be talked into calling an unfair ledger fair, and the panel says that about one woman in five eventually stops believing it.

So much for the ledger the pen can keep. Now the part it cannot.

Two clocks in one bed

Here is where my own side of the argument expects its reward, and where I have to be careful, because the body does keep time, in both sexes, and almost everything that has been sold about it is wrong.

Start with the clock that is supposed to be famous. The textbook menstrual cycle runs 28 days with ovulation on day 14, and the textbook has now been checked against 612,613 cycles logged by 124,646 users of a tracking app (Jonathan Bull, Scherwitzl and colleagues, *npj Digital Medicine*, 2019). The mean is 29.3 days with a standard deviation of 5.2. Only 13 per cent of cycles are exactly 28 days. Only 24 per cent of ovulations land on day 14 or 15. The luteal phase, the fortnight after ovulation under progesterone, is the stable part at about 12.4 days. The follicular phase before it does the varying. That is the measurable layer, and it has a consequence for the statistical one: a woman who ovulates on day 12 this month and day 20 next cannot be assigned a hormonal state by counting days on a calendar, and most of the older research that found cycle effects on the mind did exactly that. When researchers measured the hormones instead, a great deal went quiet.

Taste went quiet first, as chapter 5 told: the fertile-day preference for the square jaw did not survive hormones in the saliva. What does move is appetite. Daily saliva shows oestradiol pushing desire up and progesterone pushing it down (Roney and Simmons, 2013), and 26,000 diary entries show a mid-cycle peak in wanting sex, with the partner and otherwise (Arslan and colleagues, 2021). The volume rises. The channel does not change.

Then cognition, which is the one the self-help shelf lives on. Brigitte Leeners and colleagues tested 88 women on visuospatial working memory,

divided attention and cognitive bias at four points of a cycle, with the hormones measured, and then did it all again over a second cycle in the same women, 68 of whom completed it (*Frontiers in Behavioral Neuroscience*, 2017). In the first cycle there was a hint: working memory fell as progesterone rose from before ovulation to mid-luteal. In the second cycle it was gone. The authors called the first-cycle result what it was, a false positive, and reported no consistent link between oestradiol, progesterone or testosterone and any of the measures, between women or within them. It is the only study in the field I know of that replicated itself before publishing, and it replicated nothing.

And the body. Kelly McNulty and colleagues pooled 78 studies of exercise performance across the cycle in women not using hormonal contraception (Sports Medicine, 2020). Performance in the early follicular phase, the days of the period, was lower than in the rest of the cycle by 0.06 of a standard deviation, with a credible interval from minus 0.16 to plus 0.04, which in the notation of chapter 4 is a d so small that the two bells share about 98 per cent of their area and the difference may be zero. The authors rated the evidence low in quality and advised against population-wide training rules. For the average healthy woman, then: the cycle is measurable in the blood, visible on the thermometer, and all but invisible on the stopwatch and in the memory test. It is not the finding my side wanted, and I am giving it to you anyway.

Two things survive, and they are specific. The first is sleep. In the luteal phase progesterone raises core body temperature by about 0.4 degrees and blunts the night-time fall that sleep onset depends on. REM sleep goes down, light sleep and sleep spindles go up, and subjective sleep quality drops in the last week before the period (Baker and Driver, Sleep Medicine Reviews, 2007). That is measurable, and it is the one piece of cycle-aware living with a thermostat attached. The second is the minority for whom the cycle is not a rounding error. Pooled across 17 studies, about 48 per cent of women report premenstrual syndrome of some kind (Direkvand-Moghadam 2014), a self-report figure inflated by recall. Premenstrual dysphoric disorder, the severe and diagnosable form, affects 3 to 8 per cent, or about 5.5 per cent in the other report, and needs two cycles of daily charting to diagnose precisely because memory exaggerates. What causes it was settled by an experiment that everyone who shares a bed should know. Peter Schmidt and colleagues at the National Institute of Mental

Health in Bethesda switched off the ovaries of women with and without the disorder using leuprolide. The symptoms vanished. Then they quietly added oestradiol or progesterone back at ordinary doses. The symptoms returned, in the women with the disorder only (*New England Journal of Medicine*, 1998). In 2017, with 22 women, the group showed the last piece: it was the change in hormone level that triggered the symptoms, not the level itself, because after the first month of add-back the symptoms subsided while the hormones stayed high (Schmidt, Eisenlohr-Moul and colleagues, American Journal of Psychiatry). Women with the disorder have the same hormones as everyone else. They have a brain that responds to the rate of change, and the right drugs act within a day or two rather than the weeks an antidepressant takes for depression. Physiology, measurable, nothing a schedule can be built around, and everything a partner can stop taking personally.

Now the other clock, which my own sex prefers not to mention. Men run on a day rather than a month. Testosterone peaks in the early morning and falls through the day, and it moves with the season too (Svartberg 2003, Brambilla 2003, both in the *Journal of Clinical Endocrinology and Metabolism*), and chapter 3 has already told you how little the morning peak drives. What the daily tide does reliably is set a time of day, and that matters in a household because the two sexes' preferred hours do not agree. Christoph Randler and colleagues asked 565 people when they most wanted sex (Chronobiology International, 2014 and 2017). Women, whatever their chronotype, peaked on average between six in the evening and midnight. Men were more often ready in the morning. Only the women who were strong larks showed a second peak between six and nine, which is to say that the morning tide and the evening preference meet comfortably in a minority of beds, and the rest negotiate. Whether the women's evening peak is biology or the shape of the working day, nobody yet knows, and the report says so.

Underneath both of those clocks is a shared one, and it may matter more than either. Till Roenneberg and colleagues in Munich built a questionnaire that does not ask whether you are a morning person but when you actually sleep on days you choose, corrected for the sleep debt you ran up on days you did not (*Current Biology*, 2006). Chronotype, it turned out, is normally distributed. It drifts late through adolescence, peaking in lateness at 19.5 years in women and 21 in men, and then slowly ad-

vances with age. Roenneberg named what happens when a late chronotype meets an early school or office: social jetlag, the equivalent of flying several time zones west every Friday and back every Monday. In a couple the question is not who is right about bedtime. It is how far apart the two clocks are. Similarity in chronotype predicts relationship satisfaction, sexual satisfaction and sleep quality across dating, newlywed and older couples (Randler 2014, Sprajcer 2022, Diaz-Morales 2019). And sharing the bed itself, on the one small test of it, helps rather than hurts: 12 couples aged 18 to 29 recorded over four laboratory nights moved about one and a half times as much when together, and got about 10 per cent more REM sleep, 23 per cent of the night against 21, in longer unbroken stretches (Drews and colleagues, Frontiers in Psychiatry, 2020). Twelve couples. A direction, not a weight, but it runs against the folk wisdom about the restless partner.

What the pen can keep

So here is the tool. The second shift as a gap in total hours did not survive the diaries, and the people who told you it had were counting fifty kitchens. What survived is harder to fix and easier to see once you know where to look: not the number of minutes but their kind, the routine against the episodic, the anticipated against the decided, the interrupted against the whole. The body adds two clocks that are real. A monthly one, whose effects on mind and muscle are for most women too small to schedule and for a few too large to ignore. And a daily one that both sexes run on, together or badly. None of it is a verdict. All of it is measurable, or statistical, or ours, and the chapter has tried to say which at every step.

Keep the ledger by kind, for a week, with a column for the thinking. Then look at the clocks. Then decide, together, which lines you want to balance and which you are content to leave unequal and call fair, because the German panel says you can do either, and that it is the not deciding that costs.

The chapter has argued that the kinds of minutes carry a cost the totals hide, that the first birth is where the ledger tips, and that two bodily clocks are worth designing a shared day around. Here is the best case that it is wrong on all three, in the opponents' numbers.

On the first: the cost of "kinds" rests on thin instruments. Daminger's four stages come from 70 interviews with 35 couples. The link to maternal mental health comes from one cross-sectional survey of 322 mothers, which cannot say which way the causation runs. The multitasking gap is nine hours a week from a pager study of 693 parents in a single sample. None of it has a multinational time-use measure, and the report that supplied it lists exactly that as its open question. Set against it, the diaries: 1.2 million days, 30 countries, equal totals, and core housework falling sharply for women across four decades (Gershuny and Sullivan 2019). The strongest reading is that the second shift fell, the convergence continues, and "the mental load" is a name for the residue.

On the second: the reversion at the first birth looks like body and is script. The same German panel tradition that shows the reversion on both sides of the old border shows West German wives raising their housework when their income passes their husband's and East German wives doing nothing of the kind, across 21 years of the Socio-Economic Panel from 1991 to 2012 (Lippmann, Georgieff and Senik, *Economic Journal*, 2020). Identical physiology, identical tax code since reunification, opposite household response. If the Endocrine Accent, the body's lilt, set the ledger, the border could not. And the caregiving brain, which the chapter kept out of its argument, is the sharpest cut of all: among 89 first-time parents scanned by Eyal Abraham, Ruth Feldman and colleagues (PNAS, 2014), the network that lights up in primary-caregiving mothers lit up in primary-caregiving fathers raising infants without a mother, and a father's connectivity rose with his hours of childcare. The body follows the ledger. It does not write it.

On the third: the two clocks are one clock and a rounding error. The monthly one, measured properly, moves exercise performance by 0.06 of a standard deviation with an interval that includes zero (McNulty 2020, 78 studies) and moves working memory by nothing that survives a second cycle (Leeners 2017). Partners cannot detect fertility (Gangestad 2019, Schleifenbaum 2022). The chronotype and sex-timing data come from 565 questionnaires, the co-sleeping data from 12 couples, and the

report itself cannot say whether the female evening peak is endocrine or the shape of the working day. On that reading the chapter's physiology is the self-help shelf with error bars, and a couple who designed their week around it would be designing around the noise. The honest instrument for a household is the one the chapter started with, the pen, and the honest conclusion is the sociologists': what predicts satisfaction and sex is perceived fairness (Johnson, Galambos and Anderson 2016, 1,315 couples), which no hormone sets and no diary measures.

I cannot answer all of that, and the chapter does not try. What I can say is that the second and third objections prove the chapter's layer rule rather than defeat it: the reversion is culture, measurably, and the clocks are physiology, measurably small, and the mistake this book exists to avoid is filing either under the other's layer. The first objection stands as stated. The mental load is a real thing with a thin literature, and I have told you how thin.

IN PRACTICE

For a couple, then. In conditionals, with the places where each one fails.

If you want to know what your household actually does, keep the diary, both of you, for one ordinary week, and add the column the official booklet lacks: what you were holding in your head. Write the kind of task, not only the time. The failure mode is that the ledger becomes a scoreboard, and the panel says the scoreboard is not what predicts satisfaction, fairness is, so read it once, together, and put it away. The counter-evidence is Mikula's paradox: a couple can keep a perfectly balanced book and still feel cheated, and the reverse.

If both of you intend to keep your paid work after the first child, decide the laundry before the birth, not after, and decide who anticipates and who monitors as well as who does. The panel says couples who start symmetrical almost invariably tip at the birth and do not tip back. The failure mode is that the decision made in month seven of the pregnancy is overruled in month three of the baby by lactation, which is the one bottleneck in this chapter that no agreement moves. The counter-evidence is the eastern half of the panel, where the same bodies tip differently, which tells you the rest is negotiable.

If your division is unequal in hours and you both call it fair, it probably is, for now. The failure mode is in the panel: about one woman in five in an over-burdened household eventually stops calling it fair and starts thinking about leaving. If you are the one carrying the anticipating and the monitoring, say so in those words. Daminger's finding is that the deciding gets shared because it is visible and the rest is not shared because it is not. Make it visible: the list on the wall, the calendar both can see, the agreement that whoever notices the shoes also buys them rather than reporting them.

If you are the man and the paid hours are the thing you could cut, the German fixed-effects models say you would tend to be happier for it, and the routine work tends to grow on the people who do it. The failure mode is the other ledger. In the American panel a husband outside full-time work carried a divorce hazard roughly 30 per cent higher. The condition, then, is that both of you reject the breadwinner script before the hours change, out loud, because the script is enforced by the partner and the in-laws as much as by anyone, and the eastern half of the German panel says the script is not the body's.

If she has a cycle, and she does, do not schedule the household around it unless she asks you to. For most women the measurable effect on mind and muscle is too small to plan for, and "cycle syncing" is a product. The failure mode is that a partner who has read this chapter starts explaining her cycle to her, which is the benevolent version of the old insult and lands the same way. What does hold: the last week before the period tends to bring worse sleep and a warmer body, so a cooler bedroom in that week costs nothing and is the one piece of cycle-aware living with a replicated mechanism behind it.

If she has premenstrual dysphoric disorder, or you suspect it, treat it as physiology, because it is. Two cycles of daily charting gets the diagnosis, the right treatment acts within a day or two, and the cause is a brain that responds to hormonal change, not a character or a mood. The failure mode is on both sides: hers, in self-diagnosing from memory, which inflates the figures, and his, in reading a diagnosable disorder as a monthly excuse. The counter-evidence is the base rate: 3 to 8 per cent, so most couples arguing about the premenstrual week are arguing about something else.

If your clocks differ, measure the difference before you fight about it. A couple an hour apart in chronotype has a different problem from a couple

four hours apart, and the satisfaction data say the gap matters more than the direction. The failure mode is the working day: social jetlag is imposed by the timetable, not by the partner, and a couple cannot fix what the timetable imposes, only share it. If you can share a bed, the small evidence says do, and let the moving about happen. The counter-evidence is that it is twelve couples.

And if none of this fits, which for some couples it does not, then the overlap is doing its job. These are averages from diaries and panels. Your household is one line in the ledger, and the ledger was never going to tell you what to want. What it can tell you is where the first birth tips it, and that is the next chapter's business: the father, and the tide that goes out when he picks up the child.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- Total work equal, kinds unequal: Gershuny and Sullivan 2019, MTUS, 1.2 million diary days, 30 countries (CO-02).
- The German care gap and the reversion at the first birth: Destatis ZVE 2022 and the FReDA panel 2021 to 2023 (CO-02, SY-06).
- Cognitive labour's four stages and their allocation: Daminger 2019 (CO-02).
- The breadwinner norm's persistence in divorce risk: Killewald 2016 (CO-02, CO-06).
- Perceived fairness, not hours, predicting satisfaction and sex: Johnson, Galambos and Anderson 2016, Mikula 1998 (CO-02).
- The cycle's measured effects on cognition and performance as near zero: Leeners 2017, McNulty 2020 (CO-05).
- PMDD as sensitivity to change, not level: Schmidt 1998 and 2017 (BI-04, CO-05).
- Cycle variability: Bull 2019 (BI-04, CO-05).
- Chronotype, social jetlag and chronotype matching: Roenneberg 2006, Randler 2014 and 2017 (CO-05).

Decoration

- The woman with the pen at the cooker.
- The dripping tap.
- The twelve couples in the sleep laboratory (a direction, not a weight).
- The pension arithmetic (one country, one report).

Contested

- Whether the "mental load" is measurable beyond interviews and one survey (CO-02's own open question).
- Killewald's hazard: about 30 per cent (CO-02) or 33 per cent (CO-06).
- Kornrich 2013 against Carlson 2016 on housework and sex: era, not error, and the Kornrich figures garbled in CO-02.
- Whether the female evening desire peak is endocrine or timetable (CO-05).
- Hakim's preference theory against Frodermann 2020 (SY-06).
- Bull's user count: 124,646 (CO-05) or 124,648 (BI-04).
- PMDD prevalence: 3 to 8 per cent (BI-04) or about 5.5 (CO-05).
- Whether the luteal temperature rise justifies a cooler room: the mechanism is Baker and Driver 2007, the bedroom figure in CO-05 is sourced to a mattress retailer and is not in the text.

NUMBERS TO CARRY

1. 450 to 500 minutes of total work a day for both sexes, women about 60 per cent of the unpaid and 40 per cent of the paid: MTUS, 1.2 million diary days, 30 countries (Gershuny and Sullivan, *Annual Review of Sociology*, 2019, per CO-02).
2. Gender Care Gap 43.4 per cent, 1 hour 16 minutes a day, down from 52.4 per cent in 2012/13, and mothers of under-sixes 10.5 fewer paid hours a week: Destatis ZVE 2022, 20,000 people (per CO-02).
3. Husband not in full-time work: divorce hazard about 30 to 33 per cent higher, 3.3 against 2.5 per cent a year, post-1975 cohort, PSID 1968 to 2013, more than 6,300 couples (Killewald, *American Sociological Review*, 2016, per CO-02 and CO-06, which differ).
4. Cycle effect on exercise performance: 0.06 of a standard deviation, credible interval minus 0.16 to 0.04, 78 studies (McNulty et al., *Sports Medicine*, 2020, per CO-05).
5. Mean cycle 29.3 days, SD 5.2, 13 per cent exactly 28 days, 24 per cent of ovulations on day 14 or 15: 612,613 cycles (Bull et al., *npj Digital Medicine*, 2019, per BI-04 and CO-05).

THE LIBRARY

- CO-02 Division of labour in the modern couple: the diaries, the care gap, cognitive labour, divorce, pensions, housework and sex.
- CO-05 Cycle-aware and rhythm-aware living: the cycle's measured effects, PMDD, the male daily tide, chronotype, co-sleeping, the app data.
- BI-04 The menstrual cycle and mind: cycle variability, the ovulatory shift's fall, sleep, the PMDD mechanism, presenteeism.

- OR-04 Work design and female physiology: consulted for presenteeism and shift work, both left to chapter 11.
- SY-06 Germany, East and West: the FReDA household mechanisms and the household response to a wife's income only. The national story is chapter 13's.
- CO-06 Control-group structures: consulted for the Killewald discrepancy and the Australian census figure.

CHAPTER 8

The Father's Falling Tide

A man is lying on his back in a tube that is louder than anyone warned him, with a mirror angled above his eyes and, in the mirror, a photograph of a baby. He has been told to keep still. For a few seconds the baby is a stranger's. Then it is his own, six months old this week, and whatever moves in him when the picture changes is what the magnets are there to catch. Nine months ago, before the birth, he lay in the same tube and had his head mapped for the first time. The two maps are about to be laid over each other, and in a few places the second will turn out to have slightly less in it than the first. The scanner is in a university hospital in a city my documents do not name, so I will not either (FA-02).

He is one of 38 first-time fathers, and the study is Darby Saxbe's, published in Cerebral Cortex in 2024 (FA-02). What it found in him, and in the 37 others, is the subject of the middle of this chapter. What I want you to notice now is simpler. Something has gone out of him between the two scans, the way water goes out of a harbour in the afternoon. The harbour is still there. What has changed is how much of it you can see, and what the going-out has uncovered.

If you have had a child, or stood beside someone having one, you have lain in that tube without the tube. The Preface gave you the headline that a man's testosterone falls when he becomes a father who holds the baby. Chapter 3 gave you the numbers from Cebu as far as my reports can be trusted with them. This chapter owns the rest: what the fall is answering, what the mother's body gives up in the same months and whether "gives up" is even the right phrase, what the brain-imaging people have found

in both parents, what happens to the tenth father, and what the other clock in the house, the ovarian one and the paternal one, is doing while the couple decides whether to do this at all.

Three layers, as always. The hormones and the grey matter are measurable. What mothers and fathers do with a toddler on a staircase is statistical, two overlapping bells, and the overlap stands beside every difference in what follows. Who takes the leave is culture. The father in the tube is about to show us how the three touch.

What the hormone answers

Lee Gettler's young men in Cebu, in the Philippines, sampled in 2005 and again four and a half years later, are the study everyone cites, and chapter 3 already carried its figures as my reports give them: a median fall of 26 per cent on waking and 34 per cent in the evening, or 20.5 per cent by one account, in a sample reported as anything from 394 to 683 men (BI-02, BI-06, DH-08, chapter 3). I will not redo it. What this chapter owns is the mechanism, and the mechanism is not a leak.

Three details of the Cebu design carry the weight. The men with the highest testosterone at the start were the ones most likely to have found a partner and become fathers by the follow-up (DH-08, FA-02). The fall came after. And the men doing three hours or more of childcare a day had the lowest readings at the end (DH-08, BI-06). Read those together and the hormone is not being drained by fatherhood as a status. It is answering what the man's hands are doing. That is a physiological claim, measurable in saliva, and it is also a conditional one. The fall tends to follow the hours, so a father who is never in the room tends not to get it (BI-06). Before the child, the hormone is doing mating effort. After it, parenting effort. The two trade against each other and the body keeps the books (DH-08).

The replication I find most persuasive is not a laboratory. Among the BaYaka foragers and fisher-farmers of the Congo Basin, Gettler, Adam Boyette and colleagues reported in the American Journal of Human Biology in 2020 that the men their neighbours rated as more reliable providers and more generous sharers had lower testosterone and higher oxytocin than the less involved men (DH-08). Different continent, different economy, same direction.

Testosterone is not the only thing moving. Storey and colleagues showed in Evolution and Human Behavior in 2000 that fathers answer the cues of an infant with raised prolactin and oxytocin (FA-02). Ilanit Gordon, Ruth Feldman and colleagues followed 43 fathers through the first six months in 2010 and found the two peptides tracking different things: a father's oxytocin went with the fine-grained exchange of faces and sounds that Feldman calls affect synchrony, his prolactin with the coordinated play in which he steers the child towards the world (DH-08). And nobody has yet found a baseline difference in oxytocin between new mothers and new fathers, across several independent cohorts (FA-01). What differs is what the hormone is doing while it is there, which turns out to be the whole subject.

One network, two routes in

In 2014 Eyal Abraham, Ruth Feldman and colleagues put 89 first-time parents into a functional scanner and measured their oxytocin (FA-01, PNAS, place not given). The parents came in three kinds, by design. Mothers who were the main carer. Fathers whose partner was. And gay fathers raising an infant as a couple, with no mother in the house, who were therefore main carers by circumstance. The question was whether the parental brain belongs to a sex or to a job.

The answer came in two networks. The first is old and low, amygdala, cingulate, insula, the ventral tegmental area, the circuit that makes a cue matter: vigilance, salience, reward (FA-01). The second is cortical and newer, the superior temporal sulcus and the temporo-parietal junction, the circuit for working out what someone else wants and planning what to do about it (FA-01). In the mothers the first network dominated, and its activity rose with their oxytocin and with how closely their behaviour tracked the baby's (FA-01). In the fathers whose partner did most of the care, the second dominated (FA-01). And in the gay fathers both ran high, the amygdala as high as the mothers', the temporal sulcus as high as the other fathers', with a connection between the two that the other groups did not show (FA-01). Across every father in the study, the strength of that amygdala-to-sulcus connection rose in a straight line with the hours he spent in direct care (FA-01).

The brain did not ask who had been pregnant. It asked who was up in the night.

Each layer gets its sentence. The two networks are measurable. Which one leads is statistical and follows the role, not the sex. Who takes the role is ours.

Oxytocin follows the same rule as the networks. Same baseline in mothers and fathers, different company. In mothers it rises with affectionate touch, the singsong register, face to face (FA-01). In fathers it rises with stimulating play, with objects, with moving the infant through space (FA-01).

What pregnancy takes, and what fatherhood takes

Now the mother, and the study that changed the field's manners. Elseline Hoekzema and colleagues, in Nature Neuroscience in 2016 (BI-06 dates it 2017), scanned 25 women before they conceived for the first time and again after they gave birth, against 20 women who stayed childless, 19 first-time fathers and 17 childless men (FA-02). The mothers' grey matter had shrunk, symmetrically, in the regions that do social cognition: hippocampus and amygdala, the orbitofrontal and subgenual prefrontal cortex, the right superior temporal gyrus, the insula, the fusiform gyrus (FA-02). The change was so consistent that a classifier told every mother from every non-mother, 100 per cent (FA-02). Against the ordinary wobble between two scans of the same person, it was six times larger in the cortex and three times in the medial temporal lobe (FA-02). Behind it ran an endocrine event with no equal in adult life: oestradiol from 3.42 picograms per millilitre before conception to 12,400 before delivery, progesterone from 0.84 nanograms per millilitre to 103, then both off a cliff at birth (FA-02). The regions that had shrunk were the regions that lit when the women later looked at pictures of their own babies, and the more a woman's brain had changed, the stronger her attachment to the infant scored afterwards (FA-02).

Six years on, Magdalena Martínez-García and colleagues could still sort the same women from the controls at 91.67 per cent, and the changes still tracked attachment (FA-02, Brain Sciences, 2021). A second pregnancy changes the brain again, differently and less: the first pregnancy's reduc-

tions cover an area 79 per cent larger, with a median change of 3.1 per cent and an effect size above one standard deviation (FA-02, Hoekzema 2022).

"Grey matter lost" is the phrase the newspapers used and it needs policing. Nothing in the data says the women lost ability at anything. The authors' own word is pruning, the kind the adolescent brain does when it specialises, and the attachment correlation points the same way (FA-02, BI-06). And the mechanism is contested in the respectable way. Pilyoung Kim and colleagues, scanning 19 mothers at two to four weeks and again at three to four months after birth, found grey matter going up in the prefrontal, parietal and midbrain regions (FA-02, Behavioral Neuroscience, 2010). Eileen Luders and colleagues found increases as early as the second day (FA-02, 2020). The reconciliation most of the field now accepts, set out by Barrière and colleagues in 2021 from animal data, is two tides: a gestational ebb driven by placental steroids, then a postpartum flood driven by the baby, which rodents show as a transient swelling of the emotion and reward regions (FA-02). Contested as mechanism. Not contested as fact.

Back to the man in the tube. Saxbe, Martínez-García and colleagues followed their 38 fathers from the partner's pregnancy to six months after the birth and found what nobody had looked for because nobody thought it was there: cortical volume down and the cortex thinner across the brain, most of all in the precuneus, with an effect above half a standard deviation (FA-02). The fathers who had lost more responded more strongly to pictures of their own baby (FA-02). The fathers who had bonded most with the unborn child, who planned to take more time off, and who then spent more hours with the infant, lost the most (FA-02). So far it is the mother's story told smaller. Then the other column. Controlling for how they slept and felt before the birth, the fathers with the larger reductions reported poorer sleep, more anxiety, more distress and more symptoms of depression at six months (FA-02). The document calls it a mixed bag and for once the machine has the right phrase.

That is the falling tide, in the body rather than the harbour. What goes out with it is a layer of insulation. What is uncovered is the child, and the rocks.

Two consolations, both tentative. The first is that the going-out seems to be the price of the attunement and not a separate injury. The second comes from the UK Biobank, more than 30,000 people, in a recent PNAS

paper my document does not attribute to an author: the more children a person has parented, the higher their functional connectivity in middle and later life, in both sexes, in exactly the networks that ageing usually takes first (FA-02). If that holds, the tide comes back.

The tenth father

Now the rocks, by name.

James Paulson and Sharnail Bazemore pooled 43 studies and 28,004 participants in JAMA in 2010 and put depression in the months around a birth at 10.4 per cent of fathers, with a confidence interval from 8.5 to 12.7 (FA-02, FA-01). One father in ten. Between the third and sixth month after birth the pooled rate was 25.6 per cent, one in four, with an interval from 17.3 to 36.1 that tells you how few studies were looking in that window (FA-02). Interviews found less than questionnaires, 4.9 against 11.0 per cent, and American studies more than the rest of the world, 14.1 against 8.2 (FA-02). A father's depression and a mother's were correlated at 0.308 (FA-02, FA-01). It is one weather system over one house.

The father's version wears a different coat, which is why it goes uncounted. Depressed fathers in the literature my document summarises were about five times more likely to be working long hours away from home and five times more likely to describe themselves as isolated, with odds ratios of 5.34 and 5.28 that the document takes from a systematic review it does not name (FA-02). Irritability, aches, the office. The Edinburgh scale was written for a different patient (FA-02).

Whether it matters to the child was settled in Britain. Paul Ramchandani and the Avon cohort team, more than 10,000 children, measured fathers' depression at eight weeks and the children's behaviour at three and a half years, and found the adjusted odds of emotional and behavioural problems doubled, 2.09, and of conduct problems in boys more than doubled, 2.66, after controlling for the mother's depression and for the father's later depression (FA-01, The Lancet, 2005). The father's state at eight weeks was an independent variable in the child's life. That is the statistical layer at its most load-bearing, and the policy that follows from it is in the practical section.

The mother's version is older and more thoroughly studied, 10 to 13

per cent of mothers in rich countries and up to 20 elsewhere (FA-02), 10 to 15 per cent by another count (DH-08). Two evolutionary readings of it circulate and I carry both. Edward Hagen argued in 1999 that postpartum depression is a bargaining move, an honest and costly signal by a mother who perceives too little support or too little viability in the infant, aimed at the father and the kin (DH-08, Evolution and Human Behavior). Kardum and colleagues found, in a longitudinal cohort my document does not date, that perceived lack of family support and a low Apgar score were the strongest predictors of onset, which fits (DH-08). Jennifer Hahn-Holbrook argued in 2018 that it is mismatch instead: the exhaustion is ancestral and the relief that used to arrive with it, the aunts and the grandmother, does not (DH-08). The two readings agree on the only sentence I need. Human maternal psychology did not evolve for a flat with two adults in it (DH-08).

Which is where the cultures come in, and where the German state did something I did not expect to find in my documents. Across unrelated societies the month after birth is fenced: the Chinese zuo yuezi, sitting the month, with the mother absolved of housework and fed warm food by female relatives, the Latin American cuarentena, forty days in bed with sisters running the house (FA-02). Germany's *Wochenbett*, the childbed, is the same idea on the public purse. State-funded midwives visit the home for up to twelve weeks after a birth and look at the mother's body, the baby's weight and the pair's state of mind (FA-02, whose sources for this are an insurance blog and a thesis). I grew up in that system without knowing it was one.

And the couple, in one breath, because chapter 6 owns the couple and this chapter only owns the birth. Jean Twenge, Keith Campbell and Craig Foster's 2003 meta-analysis put parents below non-parents in marital satisfaction by 0.19 of a standard deviation, with only 38 per cent of mothers of infants reporting high satisfaction against 62 per cent of childless women (FA-02, Journal of Marriage and Family). Brian Doss and colleagues, following 218 couples for eight years, saw the drop arrive as a step at the birth (FA-02). Danielle Mitnick, Richard Heyman and Amy Smith Slep, the same year, showed childless couples sliding down a slope to the same place (FA-02). The child does not cause the fall. The child brings it forward.

What a father is for, statistically

So far the chapter has shown two bodies changing in parallel, more alike than the folklore allows. The reader who fought for Emancipation 1.0 is entitled to ask the next question with some edge. If the parental brain follows the job, what is left of "mothers and fathers as distinct contributions"? Here is what the meta-analyses leave, in their own numbers, and it is less than the shelf of fatherhood books claims and more than nothing.

Attachment first. Patty Lucassen, Marinus van IJzendoorn and colleagues pooled 95 studies and more than 3,000 children in 2011 and found infants slightly more securely attached to mothers than to fathers, by 0.07 of a standard deviation, which is to say by nothing you could see in a room (FA-01). Security to the two parents was correlated at 0.32 (FA-01). An updated meta-analysis of the parental antecedents of attachment, 174 studies and 22,914 pairs, found maternal sensitivity predicting the child's security at 0.26 and paternal sensitivity at 0.21, with overlapping confidence intervals (DH-08). Van IJzendoorn's answer to John Bowlby's single mother is the integrative hypothesis: an infant is built to hold several secure attachments at once, which in a camp insures the child against any one adult's death, and in a flat with two adults leaves it insured against neither (DH-08). In high-risk, isolated households disorganised attachment runs 0.77 of a standard deviation higher than in low-risk ones (DH-08, no primary source named).

Then the one difference that survives, in a small way. Daniel Paquette's Risky Situation puts a toddler of 12 to 18 months with its father and introduces a stranger who keeps coming closer, then a staircase, then a toy the child is not allowed (FA-01). Fathers who push the child to the edge of what it can bear and then set the limit produce what Paquette calls an activated child. Fathers who never push produce an under-activated one, who later shows more anxiety, and fathers who push without limits produce the over-activated one, who later shows more impulsivity (FA-01). A meta-analysis of 40 studies, 3,607 fathers and 3,707 mothers, puts fathers above mothers on this activation parenting by 0.27 of a standard deviation (FA-01, sourced to a table with no author named). Draw the two bells 0.27 apart and they share about 89 per cent of their area. Millions of mothers push harder than millions of fathers, as the document says in its first paragraph (FA-01). The difference is real and it is a lilt.

Rough-and-tumble play is the vehicle. Fathers wrestle, chase and throw more than mothers, most of all around the child's fourth year and more with sons than daughters, and the children who get the good version regulate their feelings more and act out less (FA-01). Susan Bögels and Vicky Phares argued in 2008 that this is what fathers are evolutionarily for: not the harbour, the sea (FA-01). A meta-analysis of parenting and childhood anxiety, 28 studies and 5,728 pairs for mothers against 12 studies and 1,019 for fathers, found that challenging parenting by the father, and not by the mother, went with less anxiety in the child (FA-01, author not named).

Language is the oddest and I like it best. Nadya Pancsofar and Lynne Vernon-Feagans, in Child Development in 2006, recorded mothers and fathers talking to their two-year-olds in rural families where both parents worked and tested the children's expressive language at three (FA-01). Mothers talked more. Fathers used more different word roots, more questions beginning with who, what, where and why, and more requests for clarification, which in practice often amounted to "Huh?" (FA-01). A father who does not know the half-word makes the child say it again. The father's vocabulary diversity at two was a unique predictor of the child's language at three after controlling for the mother's input, income and education, and a pooled analysis of 41 studies and 5,563 children confirmed a moderate correlation (FA-01, meta-analysis not named). The Bridge Hypothesis calls the father the less accommodating listener who stands between the nursery and the world (FA-01).

Now the objection, which is also in the document and which I think is half right. Joseph Pleck's argument is that the fun playmate and the demanding listener are artefacts of a man who is at home less, who defaults to high-arousal play in his short window and does not know the child's half-words because he was at work when they were coined (FA-01). The document's own open question is whether the Bridge Hypothesis survives in fathers who become the main carer, or whether their language drifts to the maternal profile (FA-01). Nobody knows. My bet, from the brain data above, is that it drifts. But a bet is not a finding, and the honest entry is that this difference lives in the behavioural layer and may belong to the cultural one.

Michael Lamb and colleagues' 1985 frame sorts fatherhood into engagement, accessibility and responsibility, which is knowing the dentist

is on Thursday (FA-01). The first two have shifted towards fathers over fifty years. The third has barely moved, and chapter 7 owns that ledger (FA-01). The brain scans measure engagement. Nobody has scanned responsibility.

One more thing belongs here, and then the clock. When the father is gone the children fare less well, and for forty years the obvious objection was that the families without fathers were also the families without money. Sara McLanahan, Laura Tach and Daniel Schneider went through the causal designs in the Annual Review of Sociology in 2013 and found a direct effect, smaller than the raw comparisons but consistent, on finishing school, on social and emotional adjustment and on adult mental health (FA-01). Colter Mitchell and colleagues found shorter telomeres in children of a cohort of 879 who had lost a father to separation, death or prison (FA-01, Pediatrics, 2017, though the document also says PNAS). And the control group that the whole argument needs exists. A meta-analysis of 54 effect sizes and 1,509 participants found the children of gay fathers adjusted within the normal range and on some measures above it (FA-01). My document attributes that analysis to Fedewa and colleagues in 2015 in its text and to Carone and colleagues in 2024 in its reference list, and I checked the PDF, which does the same, so the figure stands and the name does not. The point survives either author. What a child needs is the whole network, and the brain data say how it gets one: the second parent's amygdala learns the job (FA-01).

The other clock

Everything above happens after the decision. Most of the people this book is for are still inside it, and the decision is made against two clocks that run at different speeds in the two bodies. This is the physiological layer and I give it plainly.

A woman's ovarian reserve, the stock of follicles she is born with, was modelled by Hamish Wallace and Thomas Kelsey in PLoS ONE in 2010 from eight histological studies and 325 ovaries (FA-05). Age alone accounts for 81 per cent of the variance in that stock (FA-05). The peak is before birth, at 18 to 22 weeks of gestation, about 300,000 follicles per ovary on average and anything from 35,000 to 2,534,000 across the 95 per

cent band, a range that should be read twice (FA-05). By 30, 12 per cent of the peak remains. By 40, 3 per cent (FA-05). The population curve is rigid. The individual is not: Anne Steiner and colleagues showed in JAMA in 2017 that the blood markers of reserve predict a woman's chance of conceiving in the next months poorly, if she has no history of infertility (FA-05). Low reserve is not sterility. It is a shorter runway of unknown length.

Dik Habbema and colleagues turned the curve into a timetable in Human Reproduction in 2015 (FA-05). For a 90 per cent chance of one child without IVF, start by 32. For two children, start by 27. For three, by 23. With IVF, add three or four years: 35, 31, 28 (FA-05). Students, my document says without naming the survey, believe IVF rescues fertility well into the forties (FA-05). It does not. It buys the three years.

The man's clock is slower and runs in a different currency. Sperm quality declines measurably after 35 and conception odds fall hard only around 50, by David Dunson and colleagues' 2004 figures (FA-05). But sperm are made continuously, each round of copying with a chance of error, and Kong and colleagues showed in Nature in 2012 that the father's age is the main driver of new mutations in a child (FA-05, whose sample description for this paper, "over 1.4 million individuals", does not match what I would expect of a sequencing study). Abraham Reichenberg and colleagues then went through 2,615,081 Swedish births in the Archives of General Psychiatry in 2006, as my document has it, though the sample and the two later hazards read to me like a different and later Swedish study folded into the same citation. Against fathers aged 20 to 24, the children of fathers of 45 and older had 3.45 times the hazard of autism, with an interval from 1.62 to 7.33, 13.13 times the hazard of ADHD, and 24.70 times that of bipolar disorder (FA-05). Those are multiples of rare things and they stay rare. They are also the largest paternal effect sizes in this chapter, and they belong to the clock nobody mentions at the dinner party. The sperm-count decline, 52.4 per cent in concentration across 185 studies and 42,935 men from 1973 to 2011 by Hagai Levine and colleagues' 2017 analysis, is in chapter 3 as a contested figure and stays contested here (FA-05).

Then the behavioural and cultural layers, which decide whether the clocks are read. Henrik Kleven and colleagues' event studies of 2019 put the long-run earnings penalty of a first child at about 20 per cent for a mother in Denmark and 60 per cent in Germany, 61 by another of my doc-

uments (FA-05). A woman facing a 60 per cent penalty rationally waits, and the wait is paid in the currency above. Germany's total fertility rate sits near 1.4, Europe's at 1.4 to 1.5 against an ideal family size that Europeans themselves put at 2.0 to 2.2, a gap Tomáš Sobotka and Wolfgang Lutz measured at 0.35 children per woman in 2010 (FA-05). Martin Bujard and Harun Sulak decomposed the German fall in 2016 and found 68 per cent of it in the disappearance of third and later children (FA-05).

The strongest sentence in the fertility literature is Peter McDonald's: fertility collapses where women reach equity in university and the labour market and meet the old inequity at home (FA-05, Population and Development Review). South Korea is the demonstration, 0.72 children per woman in 2023 after roughly 280 trillion won of pronatal spending since 2006 (FA-05, which sources the spending figure to a news aggregator). Across the OECD, the best family policies buy 0.2 to 0.3 children per woman and no more (FA-05, Luci-Greulich and Thévenon 2013).

That is the tool this chapter hands over, and it is a tide table rather than a verdict. The hormones in the father fall with the hours and come back with neglect. The grey matter in both parents goes out with the attunement and uncovers the rocks. The ovarian clock is rigid in the population and unreadable in the person, the paternal clock slow and expensive. And the whole arrangement is run, in Germany, against a 60 per cent penalty and a childbed that the state still pays a midwife to visit. Measurable, statistical, ours. Now the case against.

THE STRONGEST CASE AGAINST

The strongest case against this chapter is that it has smuggled a role difference across the border as a sex difference, and the chapter's own best study is the witness.

Abraham and Feldman's 89 parents showed that the maternal network is not maternal. It is the primary carer's. The gay fathers raising infants without a mother activated the amygdala as strongly as the mothers did, and the amygdala-to-cortex connection in every father rose in a straight line with hours of care (FA-01). Saxbe's 38 fathers lost cortical volume in the same direction as Hoekzema's mothers, scaled to their hours (FA-02). Oxytocin at baseline is the same in both parents (FA-01). Attachment to

fathers is 0.07 of a standard deviation less secure than to mothers, which is zero by any sensible rounding, and paternal sensitivity predicts a child's security at 0.21 against the mother's 0.26, with intervals that overlap (FA-01, DH-08). Even the activation difference, 0.27, leaves nine-tenths of the two bells on top of each other, and Pleck's reading, that the father plays rough because he has twenty minutes and does not know the child's half-words because he was at work, has never been refuted, only answered with a question nobody has run (FA-01). On this view a chapter titled for the father's hormone has spent four thousand words showing that the hormone is a thermometer for who does the nights.

The testosterone evidence is thinner than its fame. The Cebu paper's sample is reported by my own library as 394, 465, 624 and 683, and its headline drop as 26, 34 and 20.5 per cent (chapter 3). The fall is correlated with hours of care in an observational cohort. Nobody has assigned fathers at random to nappies.

The mother's brain is contested in direction. Kim's 19 mothers gained grey matter where Hoekzema's 25 lost it, and the two-tide synthesis that reconciles them rests on rodent data and on a 2021 review rather than on a human study that measured both tides in the same women (FA-02).

The ten per cent of depressed fathers were counted by questionnaire more than by interview, 11.0 against 4.9 per cent, and in American samples more than elsewhere, 14.1 against 8.2 (FA-02), which is a range wide enough to hide a method.

And the deepest version. Rebecca Sear and Ruth Mace's 45 populations, which chapter 2 carried and chapter 9 will carry again, found the father's presence improving a child's survival in about a third of them, and the mother's mother's in most (DH-08). The Barí children with a second father survived at 80 per cent against 64 (DH-08). The ancestral record says fathers were useful, substitutable and optional, which is a strange foundation for a chapter about what a father is for. The critics of the father-absence literature add that no fixed-effects model removes every confound, and that the money a father brings has never been fully separated from the man (FA-01).

I accept most of this. The plasticity is the finding, not an objection to it, and I have said so in every section. What I keep is narrower. The second parent's brain learns the job, but it learns it with the hours, and in the countries this book is about the hours are not distributed by a coin.

The direction of the hormonal change in fathers has held across a cohort in the Philippines and a camp in the Congo. And the grey matter in 38 fathers moved with their leave. If the role is what matters, then the role is what we are arguing about, and the chapter is about who gets it.

IN PRACTICE

What follows is conditional, at the scale of a family, because that is all the evidence licenses.

If you are a father who wants the hormonal change that the headlines promise, then the Cebu and Congo data say it tends to follow the hours of direct care and not the fact of the child (DH-08, BI-06). Three hours a day is the figure in the document, which is not a prescription but a place where the effect was largest. The failure mode is expecting the softening without doing the nights, or reading its absence as a verdict on your character rather than on your timetable.

If you are the partner of a new father and you want the second amygdala in the house to learn the job, then Abraham and Feldman's straight line says the connection grows with the hours he spends alone with the baby, not beside you (FA-01). The failure mode is the gatekeeping that both sexes do without noticing: handing over, hovering, taking back. The counter-evidence is that no one has yet scanned a father who was handed the baby and left.

If you are deciding who takes leave, then the data on both parents say the one who takes it tends to be the one whose brain changes, and that the change is not free (FA-02). Saxbe's fathers who took more leave lost more cortex and slept less. The failure mode is treating the leave as a holiday and the one who takes it as the one who has it easy.

If you are a father in the third to sixth month and you find yourself at the office later than you need to be, or oddly short with everyone, then the pooled rate of depression in that window is one in four by questionnaire (FA-02), and the symptoms in men are hours and irritability more often than tears. The failure mode is that nobody is screening you, so the only screen is the person across the table. The counter-evidence is that the one-in-four rests on few studies and a wide interval.

If your partner is depressed after the birth, then the correlation of 0.308 says you are at raised risk too (FA-02), and Ramchandani's cohort

says your state at eight weeks tends to show in the child at three and a half independent of hers (FA-01). The failure mode is the division of concern in which one parent is the patient and the other is the staff.

If the mother in your house is being told she has lost brain matter, then she has, in the same regions that now light for her baby, and in the fathers too at a smaller scale, and the only study that has looked six years on says the attachment it predicted is still there (FA-02). The failure mode is the "baby brain" joke, which is a sex-difference claim dressed as sympathy, and which the fathers' data now also disprove.

If you are two adults and no one else, then the whole of the deep-history literature says you are running an arrangement no camp ever tried (DH-08), and the two evolutionary readings of maternal depression agree that the relief which was supposed to arrive does not. Build the substitute on purpose: the mother's mother if she is near, a neighbour's reciprocity if not, the state's midwife where there is one. The failure mode is pride, and its symptom is the sentence "we are managing".

If you want the child to have the father's lilt and not just his hours, then the small differences are play that goes to the edge and sets a limit, and questions that make the child say it again (FA-01). Either parent can do both. The failure mode is the father who pushes without the limit, whose children show up in the over-activated column, and the counter-evidence is Pleck: the lilt may be a side effect of absence, and a father at home all day may lose it.

If you are a couple deciding when, then Habbema's timetable is the only number that speaks to you directly: for a 90 per cent chance of two children, start by 27 without IVF or 31 with (FA-05). The failure mode is reading the population curve as your own body, which Steiner's data forbid in both directions, and the counter-evidence is the width of the band, 35,000 to 2,534,000 follicles at the start.

If the father in the plan is past 45, then Reichenberg's Swedish hazards are real and multiply small numbers (FA-05). The failure mode is panic, because a hazard ratio is not a probability. The second failure mode is the dinner-party silence, in which only the woman's age is a subject.

If you are in Germany, then you are inside the 60 per cent penalty and the twelve-week midwife at the same time (FA-05, FA-02), which is the whole paradox of this book in one household. The penalty tends to delay the child, the midwife tends to buffer the birth. The failure mode is be-

lieving that either one is nature.

And if you take nothing else from the chapter, take the direction of the arrows. The hormone follows the hands. The brain follows the hours. The role is the variable, and it is yours.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- The parental brain follows the caregiving role, not the sex: Abraham and Feldman 2014, 89 parents, connectivity linear with hours (FA-01).
- Fathers' testosterone falls with hours of care, not with the status: Gettler 2011 Cebu, Gettler and Boyette 2020 BaYaka (DH-08, BI-06).
- Pregnancy reshapes the mother's social-cognition regions, lasting and predictive of attachment: Hoekzema 2016, Martínez-García 2021 (FA-02).
- Fathers change in the same direction, scaled to engagement, with a mental-health cost: Saxbe 2024 (FA-02).
- Paternal depression at 10.4 per cent, correlated 0.308 with maternal, independently predictive for the child: Paulson and Bazemore 2010, Ramchandani 2005 (FA-02, FA-01).
- Attachment nearly equal across parents: Lucassen 2011, the 174-study sensitivity meta-analysis (FA-01, DH-08).
- The two clocks: Wallace and Kelsey 2010, Habbema 2015, Reichenberg 2006 (FA-05).

Decoration

- The man in the tube as a scene.
- The degus.
- Zuo yuezi and the cuarentena.
- The "Huh?".
- The narrator's childhood inside the Wochenbett.

Contested

- Direction of maternal brain change: Hoekzema 2016 against Kim 2010 and Luders 2020, with Barrière 2021's two-phase synthesis (FA-02).
- Whether the activation and language differences are sex or time: Paquette and Pancsofar against Pleck (FA-01).
- Whether the birth causes the couple's decline or advances it: Doss 2009 against Mitnick 2009 (FA-02).
- Gettler 2011 sample and drop, four values each (chapter 3).
- Hoekzema's year, 2016 (FA-02) against 2017 (BI-06).

- Kleven's German penalty, 60 (FA-05) against 61, with different author lists.
- The sperm decline, Levine against Boulicault (FA-05, chapter 3).
- Postpartum depression as bargaining (Hagen) or mismatch (Hahn-Holbrook) (DH-08).
- The Fedewa 2015 / Carone 2024 attribution (FA-01).

NUMBERS TO CARRY

1. 89 first-time parents: amygdala-to-cortex connectivity in fathers rises in a straight line with hours of childcare, and primary-caregiving gay fathers show both the "maternal" and the "paternal" network (Abraham, Feldman et al. 2014, PNAS, via FA-01).
2. 25 mothers scanned before and after a first pregnancy: 100 per cent classifiable by grey-matter reduction, 91.67 per cent still at six years, oestradiol 3.42 to 12,400 pg/ml (Hoekzema et al. 2016, Nature Neuroscience, and Martínez-García et al. 2021, via FA-02).
3. 38 first-time fathers: cortical volume reductions, largest in the precuneus, d above 0.5, tracking engagement and predicting poorer sleep and mood (Saxbe et al. 2024, Cerebral Cortex, via FA-02).
4. Paternal perinatal depression 10.4 per cent across 43 studies and 28,004 participants, 25.6 per cent at three to six months, correlated 0.308 with maternal depression (Paulson and Bazemore 2010, JAMA, via FA-02 and FA-01).
5. Ovarian reserve 12 per cent of peak at 30 and 3 per cent at 40 across 325 ovaries (Wallace and Kelsey 2010, PLoS ONE), and a 90 per cent chance of two children means starting by 27, or 31 with IVF (Habbema et al. 2015, Human Reproduction), both via FA-05.

THE LIBRARY

FA-01: mothers and fathers as distinct contributions: Abraham and Feldman, Lucassen, Paquette, Bögels and Phares, Pancsofar and Vernon-Feagans, Lamb, McLanahan, Mitchell, Paulson, Ramchandani, Fedewa/Carone.

FA-02: the postpartum period and the couple: Hoekzema, Martínez-García, Kim, Luders, Saxbe, Strathearn, Storey, Twenge, Doss, Mitnick, Paulson, the Wochenbett, zuo yuezi, breastsleeping.

FA-05: fertility and the modern life course: Wallace and Kelsey, Steiner, Habbema, Dunson, Kong, Reichenberg, Levine, Kleven, Sobotka and Lutz, Bujard and Sulak, McDonald, Luci-Greulich and Thévenon, South Korea.

BI-06: the endocrine life course: Hoekzema (dated 2017), Gettler with the three-hour figure and N 683, the oxytocin and prolactin rise in fathers.

DH-08: ancestral child-rearing: Gettler 2011 with N 624, Gettler and Boyette 2020, Gordon and Feldman 2010, the 174-study sensitivity meta-analysis, the integrative hypothesis, Hagen, Hahn-Holbrook, Sear and Mace, the Barí.

CHAPTER 9

Who Keeps Children Alive

A woman of about seventy is standing at a sink in a kitchen that is not hers. It is twenty to five on a weekday afternoon, in a third-floor flat in a German city, and the child she collected from nursery at four is on the floor behind her, working a wooden train through the legs of a chair. She still has her coat on. She knows where the plasters are kept in this flat and where the spare key is, and she has a view on the contents of the fridge that she mostly keeps to herself. Her daughter's train gets in at twenty past five. Her son-in-law is in another city until Thursday. Nobody has written any of this down. There is no rota on the fridge with her name on it, and if you asked her whether she was providing childcare she would look at you the way she would look at someone asking whether she was providing air.

Stand in the doorway a moment and watch whose hands do what.

The child falls, the train hits the skirting board, and the hands that pick him up are twenty years past their last baby and not noticeably slower for it. Tomorrow's lunch is already in a box on the middle shelf. When her daughter comes in at half past five there is a handover of perhaps four sentences, two of them about the child's digestion, and then the older woman takes the tram home, and what happened in this flat between four and half past five appears in no statistic the city collects, because the city does not count it.

The Preface left you with a killer whale and a sentence I asked you to hold: that a body which switches off its own reproduction and then lives another thirty years is a body that is for something other than the next child. It called that the *Grandmother Clause*. Chapter 1 found the clause

in the whales. Chapter 2 found it at the campfire. This chapter brings it indoors.

The woman at the sink is the clause, enforced at the scale of a flat. Nobody drafted her into it, nobody can make her do it, and nobody, as we are about to see, can be sure that her doing it is good for everyone in the room, because the record on grandmothers is not a sentimental one. It is a record of who lived and who did not, kept in parish books, pension rolls and the growth charts of children, and it contains both the most reliable helper the human child has ever had and the relative that one eighteenth-century dataset names, in the title of the paper, as the devil in the house.

The question of this chapter, which is also the title of the paper the chapter stands on, is who keeps children alive. The answer is not the one my mother's generation was given. It is not the one their mothers were given either.

Forty-five populations

In 2008 two anthropological demographers working in London, Rebecca Sear and Ruth Mace, published a paper in *Evolution and Human Behavior* with a title that reads like a child's question: who keeps children alive? The method was plain. Find every population, historical or living, in which births and deaths were recorded well enough for a child's chance of surviving to be matched against which relatives were alive and nearby, and ask of each relative whether their presence made a measurable difference. They found 45 such populations, from parish registers to anthropologists' field notes, all of them without contraception and mostly without doctors. Chapter 2 carried the shape of the result and left the arithmetic to this chapter. Here it is.

The mother mattered in every population where her presence was tested: 100 per cent. Her death was the single most reliable predictor of her child's. Older siblings improved a younger child's survival in 83 per cent of the studies that looked. The mother's mother improved it in 69 per cent. And the father, the figure around whom my mother's generation's whole argument about breadwinning was built, improved his child's survival in 32 per cent of the populations studied: two of four hunter-gatherer groups, five of eighteen horticultural ones. In the other two thirds his

presence or absence made no measurable difference to whether the child lived.

Let me say which layer that lives in, because it is easy to hear wrong. It is statistical, and it is statistical across populations, not across fathers. It does not say that a father is unimportant to his child. It says that in most of the pre-modern world, if a father died, the child's odds of surviving did not move, because other hands closed over the gap, and that if the mother's mother died, they did. A finding about which deaths a household could absorb is not a finding about who loved whom. It is a finding about which hands the human household was built around, and the number of those hands was not two.

Those hands, in the forager camps of chapter 2, belonged to a dozen people a day, and I am not going to repeat the Efe or the Aka here. What this chapter adds is what the record says when the camp becomes a farm, a parish, a pension roll and a flat. The answer, it turns out, depends on whose mother the grandmother is. Sear and Mace's 69 per cent for the maternal grandmother sits beside a result for the paternal grandmother that is not a percentage, because it does not point one way. Her effect was variable, and in some populations it was negative. The next section is about those populations, and the reader who is fond of grandmothers should brace.

First the physiology underneath, briefly, since chapter 1 owns the deep time. Human women live decades past their last birth, which almost no other mammal on land has thought worth doing, and three explanations compete for why. Kristen Hawkes's grandmother hypothesis says the old woman fed her daughter's weaned children so the daughter could bear again sooner, and chapter 2 has her at the digging stick. The mother hypothesis, from Samuel Pavard and colleagues in 2008, says that a woman who bears late risks dying in childbirth and orphaning the children she already has, so the body stops as a stop-loss. And the reproductive conflict hypothesis, Michael Cant and Rufus Johnstone in *PNAS* in 2008, starts from the fact that in our species it was mostly the woman who moved to her husband's people at marriage. A young wife arriving in the house shares no genes with her mother-in-law, while the mother-in-law shares a quarter of hers with the young wife's children. If both breed at once they compete for the same food, and the arithmetic is lopsided: the older woman loses more by competing than by yielding, so she yields, and

over enough generations the yielding became menopause. I don't know which of the three is right, and the people who do this for a living have not agreed either. Notice what the third predicts, though. It predicts that the grandmother who yields is the husband's mother, and that she and her daughter-in-law are rivals by construction.

The devil in the house

Krummhörn is a stretch of farmland in East Frisia, in the far northwest of Germany. Its parishes kept registers of births, marriages and deaths through the eighteenth and nineteenth centuries complete enough for families to be reconstructed across generations, and Eckart Voland and Jan Beise went through about 32,000 of those records to ask Sear and Mace's question one grandmother at a time. A maternal grandmother alive and nearby lowered infant mortality. A paternal grandmother alive in the household raised the chance that a newborn died in its first month by a factor of 2.48. The paper is from 2002, and the title the same two authors gave to their work on these registers, which I did not write, is "The husband's mother is the devil in house".

That is one parish. Here is a village in Mali. The Dogon are patrilocal, which means a bride moves to her husband's compound, and Beverly Strassmann, an anthropologist at the University of Michigan, followed their children for years and reported in *PNAS* in 2011 that a child under five living with a paternal grandmother faced twice the hazard of death, a ratio of 2.0. Dogon girls, she found, grew faster when their maternal grandmother was absent, and her reading is that the girls were spared the weeding they would otherwise have done for the old woman. Her larger argument is the one the grandmother hypothesis has to answer: in a household where calories are short, an old woman is not a helper but another mouth.

Here is a third country. Simon Chapman and colleagues took 32,000 people from the historical parish records of Finland and tested 36 different kinds of relative against child mortality, in *Current Biology* in 2019 and the *Proceedings of the Royal Society B* in 2023. A living maternal grandmother cut a child's risk of dying from smallpox by about 30 per cent and from diarrhoeal disease by about 40 per cent, hazard ratios of 0.70 and 0.60. Pa-

ternal grandmothers and paternal cousins raised the risk. And the lethal effect of the father's side was strongest in the wealthiest families, which is not what a story about scarcity predicts. It is what a story about inheritance predicts. Where there was most to pass on, the husband's mother's presence was most dangerous to the children of the woman who had married in.

Three populations, three continents, the same sign. Now the mechanism, or one candidate for it. Harald Euler and Barbara Weitzel asked 1,857 adults in 1996 how much care each of their four grandparents had given them, and found an order that did not vary: maternal grandmother most, then maternal grandfather, then paternal grandmother, then paternal grandfather last. The old explanation was paternity. A mother's mother knows the child is hers. A father's father is unsure twice over. In 2010 Molly Fox and colleagues proposed something stranger, in a paper whose title accuses grandmothers of playing favourites. A grandparent shares a quarter of their ordinary chromosomes with any grandchild. The X chromosome is not ordinary. A paternal grandmother gives her X to her son, and her son gives it to every one of his daughters and to none of his sons. So she shares half of her X genes with a granddaughter and none at all with a grandson. Across seven historical populations, Fox found that a paternal grandmother's presence raised a granddaughter's survival and lowered a grandson's. The favouritism is in the chromosome before it is in the kitchen.

I want to be exact about the layers here, because this is the kind of finding that gets carried out of the room as a rule. The relatedness is measurable: it is arithmetic. What it does to a child's odds is statistical, and the sign flips with where the household is and what it owns. What a family does with the news is cultural, and the paternal grandmother at the sink in the opening scene, if she is one, is not statistically more dangerous to anybody in a city with vaccines and a supermarket, because the mechanism she was dangerous through was food and inheritance, not malice. The devil in the house was a hungry house.

The pension and the granddaughter

If the grandmother's effect runs through food, then give her money and watch where it goes. That is what South Africa did, by accident, in the early 1990s, and what Esther Duflo, the economist at MIT, did on purpose in a paper in the *World Bank Economic Review* in 2003.

The country extended its old-age pension to Black households at a level worth roughly twice the median rural income per head, paid to the old person and not to the household. Duflo asked what happened to the bodies of the children who lived with the pensioner. She measured weight-for-height and height-for-age, the two numbers a clinic uses to tell whether a child is being fed. When the pension went to a grandmother, her granddaughters gained 1.19 standard deviations in weight-for-height and 1.16 in height-for-age. Her grandsons gained nothing measurable. When the pension went to a grandfather, nobody gained anything, granddaughter or grandson. The same money, the same house, the same children, and the answer to "did the child grow" depends on the sex of the hand that holds the cash and the sex of the child. Economists had modelled the household as a single purse. Duflo's finding is that it is not a purse. It is a set of relationships with money passing through them, and the relationships have sexes.

Now us. The grandmother who kept children alive in 45 populations did it in conditions of scarcity, and most of my readers do not live in scarcity. So what does the clause do in a country with vaccines? Nobody has measured it as well as Duflo measured the pension, and the evidence has two halves that do not agree.

The first half says the household still runs on more hands than two, and suffers when it has only two. Moïra Mikolajczak and Isabelle Roskam and colleagues surveyed parents in 42 countries for the International Investigation of Parental Burnout, published in 2020 and 2021, and found that the strongest predictor of parental burnout at the level of countries was not wealth, not inequality and not how much the mother worked, but how individualist the culture was. Where the extended family is still expected to turn up, parents burn out less, even when everything else is harder. A review of cohousing in rich countries, 25 studies and 77 communities, found more mutual support and less isolation, and in one study of mothers' neighbourhoods, low connectedness carried an odds ratio of

1.50 for physically harsh parenting. Hands buffer tempers.

The second half says the hands have to be chosen. Sear's own 2016 review of kin and fertility found that living near relatives raised fertility in historical populations, including 3.1 million couples in the United States census of 1900 to 1910, and that in modern high-income societies the effect of kin is mixed or negative. A study of young adults forced back to their parents' homes by rents and insecure work found their mental health 4.03 points lower on a 100-point scale, with a confidence interval from 5.93 to 2.14. The extended household, in other words, is a buffer when it is voluntary and a cage when it is not, and the camp of chapter 2 was voluntary in a sense that a three-bedroom flat with a mother-in-law in the box room is not. The Agta of the Philippines, whose children's carers Abigail Page tracked with proximity sensors across 1,701 child-carer pairs in 2019, give care to kin in proportion to the child's need and to non-kin in proportion to what the other household gives back. Reciprocity predicted care as strongly as blood did. Nobody in that camp was trapped with anybody.

Two more numbers belong here because they are about the father, whom this chapter has so far treated unkindly. Marinus van IJzendoorn and colleagues pooled 174 studies and 22,914 parent-child pairs on what predicts a secure attachment, and found that a mother's sensitivity correlated with her child's security at r = .26 and a father's at r = .21, with overlapping confidence intervals, which is to say equivalently. A child does not check the sex of the hands. Van IJzendoorn's resolution of what he calls the multiple caretaker paradox is that the infant is built to form a network of attachments, not one, which is what you would expect of an animal raised by a dozen people a day. Where the network collapses, in families under severe economic and social strain, disorganised attachment rises with an effect size of 0.77, which is large by the standards of this book. Children, it appears, were never designed for two carers, and are worst off with one.

The other number is the daycare one, and it is smaller than it sounds. Harriet Vermeer and van IJzendoorn pooled nine studies and 303 children in 2006 and found that cortisol, which should fall across the day, rose from morning to afternoon at centre-based nursery and did not at home. The effect was r = .18, with a confidence interval from .06 to .29, and strongest under 36 months. That is measurable, it has replicated, and it is modest. What it means for the child in later life is not known, and

the study my mother's generation would have wanted, comparing a nursery to an isolated flat with one exhausted adult rather than to an idealised home, has not been run. Across the hall, the Aka and Ngandu children Adam Boyette watched in 2016, 50 and 48 of them, were taught by adults only until the age of four or five and by other children after that. Adults leave the frame early in most of the human record. That is not an argument for any particular nursery. It is a reason not to assume that the two-adult flat is the baseline the nursery is departing from.

Trucks, dolls and the parents on camera

Now the children themselves, and the question my mother's generation thought it had closed: whether a boy and a girl arrive different or are made different. Three layers, in order.

The measurable layer starts with the girls chapter 3 introduced, those born with congenital adrenal hyperplasia, exposed to adrenal androgens from about the seventh week of gestation, who as children choose trucks and blocks more than their sisters do. Chapter 3 has Vickie Pasterski's 2005 numbers. What belongs here is what her cameras saw. The parents in that playroom were coded, and they praised their daughters with the condition for touching dolls more than they praised their unaffected daughters, which is the opposite of the socialisation story, and the girls went on choosing the trucks, including when they played alone with nobody to please. The more severe the condition, the stronger the preference. That is a dose-response curve in a human child, and it is as close to an experiment as ethics allows.

The monkeys do the rest. Janice Hassett, Alexander Siebert and Kim Wallen put wheeled toys and plush toys in front of a troop of 135 rhesus macaques in 2008 and analysed 34 juveniles. The young males went to the wheeled toys with a consistency the females did not show. The young females played with both. Gerianne Alexander and Melissa Hines had found the same in 63 vervet monkeys in 2002, and in 2009 Alexander tracked the eyes of 30 human infants aged three to eight months and found the girls looking at a doll over a truck at an effect size above 1.0 and the boys looking at the truck at 0.78. By the time Brenda Todd and colleagues pooled 113 effect sizes from 75 studies in 2018, the preference had

the magnitude of a physical difference: boys for boy-typical toys at d = 1.83, girls for girl-typical toys at 1.60, detectable by twelve months. Those are among the largest behavioural sex differences anybody has measured, in the same league as height, which chapter 4 used as its yardstick, and like height they overlap. The girl with the truck and the boy with the pram are in every nursery, and nothing in a bell curve tells you which child in front of you they are.

And here is the thing the measurable layer cannot yet explain. At birth there is nothing. Simon Baron-Cohen's group reported in 2000 that newborn boys under a day old looked longer at a mobile and newborn girls longer at a face, 102 babies, and the finding went round the world. A meta-analysis of 31 studies, 40 experiments and 1,936 newborns found no difference in how long the sexes looked at faces, an effect of 0.076 and a p-value of 0.321. Both sexes prefer faces. So: nothing at one day, a large difference at twelve months, and the field does not know what happens in between. The report I am working from lists that as an open question, and I am going to leave it open rather than fill it with a story.

The cultural layer has a surprise in it. The usual account of the toy aisle is that parents make it: boys get trucks because fathers hand them trucks. Joyce Endendijk and colleagues tested that in 2016 by pooling 126 observational studies of 15,034 families, observed and not self-reported, on two things parents demonstrably do: control their children and support their autonomy. Parents were slightly more controlling with boys, at d = 0.08, and treated the sexes identically on autonomy, d = 0.03. In the post-industrial world, on the dimensions that were measured, parents treat sons and daughters almost exactly alike. That does not mean culture is absent. It means that whatever is sorting the nursery into trucks and dolls, it is not mainly the parents' hands, and the report's reading, which I find plausible and cannot prove, is that children sort themselves, choosing the activities, the playmates and the models that fit a bias they arrived with. Marketing then sells to the sort.

Between the two is the behavioural layer, and it has one more finding this chapter needs. Chapter 4 gave you the childhood temperament meta-analysis of Nancy Else-Quest, Janet Hyde and colleagues from 2006 at its trait level: girls ahead on inhibitory control, boys on activity. The same paper's overall factors, in the older cohorts, are larger. Effortful control, the capacity to sit still, wait, and do the dull thing first, favours girls at d

= 1.01. Surgency, the appetite for speed, noise and high-intensity pleasure, favours boys at 0.55. Negative affectivity, which is sadness and fear, shows no difference at all: 0.06. The gap in self-regulation widens through childhood and is at its widest in the years when a child is asked to sit at a desk for six hours a day. That is a statistical fact about populations, with the overlap that every *d* in this book carries, and the next section is about what a culture built on it.

A school built for one of them

Daniel and Susan Voyer pooled 502 effect sizes from 369 samples in the *Psychological Bulletin* in 2014, every one of them a teacher's mark, from primary school to university. Girls were ahead in every subject, at a mean of d = 0.225, most in languages at 0.374 and least in mathematics at 0.069, which was still a lead. And the finding that matters is the one about time. Girls have always been ahead in teachers' marks, as far back as the samples go. The boy crisis is not a crisis in the grades. It is a crisis in what the grades are now worth.

That lead is small on the scale chapter 4 gave you, close to the 0.20 for which the two bells share 92 per cent of their area. What makes it bite is what it is a lead in. A teacher's mark rewards exactly the thing the temperament meta-analysis says girls arrive with more of: effortful control, the dull thing done first, the homework in on time. A standardised test does not, and on standardised tests the boys historically held a small lead in mathematics and science. Then came the pandemic, and the 2025 round of the international PISA tests, which the OECD reported as the steepest declines in the test's history. Reading scores for boys fell by 17 points on the OECD average against 12 for girls. Girls were ahead in reading in 89 of 90 participating systems, and 37 per cent of boys failed to reach the baseline level in reading against 25 per cent of girls. In mathematics boys kept a slight average lead, carried by a small group at the top of the distribution, while the median boy struggled. Richard Reeves of the Brookings Institution put the structural reading of this in *Of Boys and Men* in 2022: boys' prefrontal cortices mature later, so a boy and a girl of the same age in the same classroom are at different stages, and the classroom was built for one of them. His remedy, starting boys a year later, is a policy nobody

has tested at scale, and it is listed in my sources as an open question, not a finding.

The remedy that has been tested is the one the Preface listed among the fallen, and it is time to say why it fell. Erin Pahlke, Janet Hyde and Carlie Allison's 2014 meta-analysis of single-sex schooling, 184 studies, 1.66 million pupils, 21 countries, found what the brochures promise when it compared schools raw, and then asked who was in them. Single-sex schools are disproportionately selective, private and well-off. Control for prior achievement and family background and the advantage is g = 0.10 for girls and 0.06 for boys in mathematics, 0.06 and 0.04 in science. A tenth of a standard deviation is the size of a good night's sleep. The argument has not stopped, which is why the finding is worth deepening. Leonard Sax and Michael Gurian, the movement's advocates, hold that boys and girls need different room temperatures, different lighting, confrontation for one and collaboration for the other. Hyde and Diane Halpern answer that the overlap in the two sexes' brains is vast, that separating children teaches them the stereotype, and that a school which never puts a girl and a boy on the same project is poor preparation for every workplace they are going to enter. The meta-analysis cannot measure the harm either side alleges. It can measure the benefit, and the benefit is a rounding error.

What the evidence does support is quieter and cheaper than a separate school. Thomas Dee matched pupils to teachers in the *Journal of Human Resources* in 2007 and found that a boy assigned to a male teacher did better, was rated better and was more engaged, enough to close up to a third of the gap in English, with a mirror effect for girls in science and mathematics with a female teacher. David DuBois and colleagues pooled 55 to 70 evaluations of youth mentoring programmes, more than 25,000 young people, in 2011, and found modest effects, d = 0.18 to 0.21, that depended entirely on duration: a mentor who stayed more than a year reduced behaviour problems and raised self-worth, and a mentor who left inside six months did measurable harm. Neither works by teaching differently. They work by showing a boy that a man reads, and by not leaving.

Then adolescence, where the two sexes lose different things. Carol Gilligan's 1982 book *In a Different Voice* documented what she called a loss of voice in adolescent girls, the moment when a girl who argued at eleven learns at fourteen that keeping her relationships means keeping her objections to herself, and later work, Theran in 2003, tied that silence to the

rise in girls' depression and eating disorders in the same years. Niobe Way, a developmental psychologist at New York University, spent hundreds of longitudinal interviews on the boys and published them as *Deep Secrets* in 2011. The boys at thirteen and fourteen told her they loved their best friends, told them their secrets, said they would go to pieces without them. Their descriptions, she writes, read less like boys stranded on an island and more like a love story. By seventeen the same boys had stopped. They had learned that the intimacy was readable as feminine or as gay, and they put on what Way calls a mask, and the loss of the friendships tracks the rise in depression, isolation and suicide among young men. Her claim is that this is not a defect in boys. It is a crisis of connection, enforced from outside, at a known age. And Gilligan, revising her book in 2023 under the title *In a Human Voice*, concedes the point from the other direction: the ethic of care she once called a different voice is not female at all, she now says, but human, coded as feminine by a culture and then taken from boys.

Here, then, is the tool this chapter hands over. At home the question is not whether boys and girls differ, which they do, not at all on the first day, a lot by the first birthday and most of all in the capacity to sit still at nine. Nor is it whether the parents made the difference, which on the best evidence they mostly did not. The question is which hands are in the room, whose mother they belong to, what the room costs, and what the room does with a child who is at the far end of one of the bells. The woman at the sink in the opening scene is the oldest answer to the first of those questions. The parish books say she is also, under the wrong conditions, the oldest problem. The clause does not say she must be there. It says the household was never built without someone like her, and that the modern flat is the first one that has tried.

THE STRONGEST CASE AGAINST

Take the chapter's two halves at their weakest, in the voice of the people who have made the argument.

The grandmother first. Beverly Strassmann's reading of her own Dogon data in *PNAS* in 2011 is that the grandmother hypothesis has been sold on the populations where it happened to work. Where food is short and the household is patrilocal, the old woman is a competitor for calories, her co-residence doubles a child's hazard of death, and girls grow

faster when she is absent. Chapman's Finnish records put the paternal side's lethal effect highest among the rich, which means it is not even a scarcity story but an inheritance story, and inheritance is culture. Fox's X-chromosome result is a seven-population correlation with a mechanism nobody has observed operating. Sear's 2016 review says that in high-income societies kin have mixed or negative effects, and the forced co-residence data, a 4.03-point loss of mental health, say that putting the generations back under one roof in a modern economy makes adults worse off. In short: the grandmother who kept children alive did so in a world of smallpox and short rations, and the clause this chapter wants to carry into a flat with a fridge is a clause from a lease that expired. What the modern household needs is hands. Van IJzendoorn's 174 studies show that the sex of the hands does not matter to the child, r = .26 for mothers and .21 for fathers, and the Agta data show that need and reciprocity, not blood, predict who gives care. The Grandmother Clause, at home, reduces to "get help", and the word grandmother in it is sentiment.

The children second. The chapter leaned on toy preferences at d = 1.83 and 1.60 and on temperament at 1.01 and 0.55. Lise Eliot's 2021 synthesis of three decades of human brain imaging, after correcting for the difference in size, leaves sex explaining about 1 per cent of the variance in brain structure, and the task-based imaging that was supposed to show sex-specific circuits has not replicated. The newborn face-and-mobile study, 102 babies, fell to 1,936. Endendijk's 15,034 families show that parents treat the sexes alike, which the chapter read as evidence that the difference is in the child, but which can be read with equal force the other way: the differences that remain are produced by the socialising agents the meta-analysis did not measure, which are peers, screens and shops. Gilligan in 2023 says the ethic of care is human. Way says the boys' isolation is enforced. The grade gap of 0.225 is a fifth of a standard deviation, which chapter 4 taught you to read as 92 per cent shared, and the single-sex schools built on the difference deliver g = 0.10 for girls and 0.06 for boys when you ask who was in the room. Alex Thornton and Katherine McAuliffe argued in 2015 that cooperative breeding does not even buy the cognition that Sarah Hrdy credits it with, since the monkeys that breed cooperatively are no cleverer than their cousins. On this reading there is no family-scale sex difference worth designing around. There are two small bells, one measured in trucks, and a culture that turns small gaps into separate schools and

lonely boys.

I think that case is right about the flat with a fridge and wrong about the hands. The chromosome arithmetic is a correlation, agreed. Duflo's pension is not. The same money, in a grandmother's hand, grew granddaughters by more than a standard deviation and grandsons not at all, and in a grandfather's hand grew nobody, and the ration in that house was not short. Whatever filters help through the sexes of the giver and the child was still working in 1993 with cash. That is the finding the case against has to explain, and it has not.

IN PRACTICE

This is the family's scale, so the you here is a household: two adults, or one, and the children, and whoever else is within reach of the door.

If the mother's mother is within an hour's travel, then the oldest arrangement in the record is available to you, and the panel data say it tends to lower parental burnout and raise the number of hands at the end of the day. The failure mode is the one Sear found in rich countries: kin whose help comes with a view on how the child should be raised, so that each hour of relief costs an hour of argument. The counter-evidence is that proximity is not co-residence, and that the Agta kept their helpers by reciprocity, not by obligation. Ask her what she wants back.

If the help on offer is the father's mother, then the parish books are not a verdict on her, because the mechanism they measured was food and inheritance and your fridge is full. The failure mode is still there, in a different currency: the reproductive conflict model predicts friction between a woman and her mother-in-law by construction, and the Finnish data put it highest where there was most to inherit. Name the currency before she moves in. If she is to move in, then the co-residence data say it tends to go badly when it is forced by money, and the cohousing data say it tends to go well when each party keeps a door that closes.

If you are two adults and nobody else, then you are running an arrangement that appears nowhere in 45 populations, and the survey of 42 countries says that the individualism which makes it possible is the strongest predictor of exhaustion in it. The failure mode is to treat the exhaustion as a defect in the marriage or in one of you. The counter-evidence is the

Agta: unrelated households that traded childcare did as well as kin. Find a second household and keep accounts, lightly.

If the child is under three and the choice is a nursery or one adult alone in a flat all day, then the cortisol data, r = .18 across 303 children, say the nursery tends to raise an afternoon stress hormone, and say nothing at all about the flat, because nobody has measured the flat. The failure mode is to read a small, real physiological effect as a verdict on working mothers, which is exactly what the language of my mother's country stood ready to do. The counter-evidence is the forager record: adults leave the frame at four or five, and children have been raised in groups of other children for most of the time there have been children.

If you have a son who cannot sit still, then the temperament data put him at the wide end of a bell that is wider for boys, d = 0.55 on surgency in the older cohorts. The failure mode is to treat the play as the problem. The counter-evidence is Endendijk: parents already control boys only slightly more than girls, d = 0.08, so a decision to let him wrestle is not a decision to raise him differently. It is a decision to let him be where he already is.

If you have a daughter who chooses the truck or a son who chooses the pram, then the toy data at d = 1.83 and 1.60 say that the averages are far apart and that your child is one of the many who are not at them. The failure mode is correction. Pasterski's parents corrected, on camera, and the girls went on choosing the trucks when they played alone. The counter-evidence is only that the average is real, which it is, and which tells you nothing about the child on your floor.

If your boy is behind at school at six and the teacher's marks say so, then the marks measure effortful control before they measure anything else, girls lead on it by about a standard deviation in the older cohorts, and the lead has been there in every sample since marks were kept. The failure mode is the separate school, which delivers 0.06 of a standard deviation after you ask who else attends. The counter-evidence is Dee: a male teacher closed up to a third of the English gap in one matched study, and a mentor who stayed a year did more than one who left in six months, who did less than no mentor at all. Find him a man who reads and does not leave. If the school offers to start him a year later, then Reeves's case is reasonable and untested, and you are the test.

If your son is fifteen and his friends have thinned to nothing, then Way's interviews say the thinning tends to happen at that age, is enforced

from outside, and tracks the rise in young men's depression. The failure mode is to read his silence as the male condition. It was not his condition at thirteen. The counter-evidence is that Way's boys were interviewed, not followed into adulthood with controls. Ask about the friend by name. If your daughter is fourteen and has stopped arguing with you, then Gilligan's reading is that she has learned the price of a voice, and the failure mode is relief. The voice she loses, on Gilligan's own revision, is not a girl's voice but a human one, and the boy down the hall is losing the same thing by another route.

If the father is the one at the sink, then the attachment data say the child does not know the difference, r = .21 against .26 with overlapping intervals, and chapter 8 has what his hormones are doing about it. The failure mode is the one the 45 populations record: that the household's survival never depended on him, and so nobody built the hours around him. The counter-evidence is that the populations also never had a fridge, and that what the household depends on now is time, which he has as much of as anyone.

And if you are the woman at the sink, then nothing in this chapter tells you what to want. It tells you that the flat was built on the assumption of your hands, that no one in it has counted them, and that the clause was always written in your favour and never signed by you.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. Sear and Mace 2008, 45 populations: mother 100, siblings 83, maternal grandmother 69, father 32 (FA-03, DH-08). The paternal grandmother's negative sign in Krummhörn (Voland and Beise 2002), the Dogon (Strassmann 2011) and Finland (Chapman 2019, 2023) (FA-03). Duflo 2003, the pension that grew granddaughters and not grandsons, in grandmothers' hands only (FA-03). Van IJzendoorn's equivalence of maternal and paternal sensitivity, 174 studies (DH-08). Endendijk 2016, parents treat the sexes alike, d = 0.08 and 0.03 (AN-10). Todd 2018 toy preferences at d = 1.83 and 1.60 and the Pasterski cameras (FA-04, AN-10). Voyer and Voyer 2014, the stable grade lead at 0.225 (FA-04). Pahlke 2014, why the single-sex finding fell: selection (FA-04).

Decoration. The woman at the sink. The vervets and the rhesus troop. Reeves's redshirting proposal. The devil-in-the-house title. The good night's sleep.

Contested. Why menopause exists: Hawkes against Pavard against Cant and Johnstone (FA-03). Whether the paternal grandmother's effect is X-chromosome favouritism, calorie competition or patrilocal power (FA-03, its own open question).

Whether the daycare cortisol rise means anything for outcomes (DH-08). Whether cooperative breeding bought cognition (Hrdy and Burkart against Thornton and McAuliffe) (DH-08). What happens between the newborn's null and the one-year-old's d above 1.0 (AN-10, its own open question). Whether Endendijk's null on parenting means the difference is in the child or in the agents not measured (AN-10). The PISA 2025 figures, pending the primary volume (FA-04).

NUMBERS TO CARRY

1. 45 natural-fertility populations: the mother's presence improved child survival in 100 per cent of studies, older siblings' in 83, the maternal grandmother's in 69, the father's in 32 (Sear and Mace 2008, *Evolution and Human Behavior*, per FA-03 and DH-08).
2. Krummhörn, about 32,000 parish records, eighteenth and nineteenth centuries: a paternal grandmother in the household raised neonatal mortality 2.48 times, a maternal grandmother lowered infant mortality (Voland and Beise 2002, per FA-03, which names the journal two ways).
3. South Africa, early 1990s: a pension to a grandmother raised granddaughters' weight-for-height by 1.19 standard deviations and height-for-age by 1.16, with no effect on grandsons and none from grandfathers (Duflo 2003, *World Bank Economic Review*, per FA-03).
4. Toy preference: boys for boy-typical toys d = 1.83, girls for girl-typical toys d = 1.60, 113 effect sizes, 75 studies (Todd et al. 2018, per the FA-04 PDF). Parents' control of boys versus girls: d = 0.08, autonomy support 0.03, 126 studies, 15,034 families (Endendijk et al. 2016, *PLoS ONE*, per AN-10).
5. Single-sex schooling after controls: g = 0.10 for girls and 0.06 for boys in mathematics, 0.06 and 0.04 in science, 184 studies, 1.66 million pupils, 21 countries (Pahlke, Hyde and Allison 2014, *Psychological Bulletin*, per the FA-04 PDF). Teachers' marks: girls ahead at d = 0.225 overall, 0.374 in languages, 0.069 in mathematics, 502 effect sizes (Voyer and Voyer 2014, per the FA-04 PDF).

THE LIBRARY

- FA-03 Grandmothers, alloparents and the extended household: Sear and Mace, the three menopause hypotheses, Euler and Weitzel, Fox, Strassmann, Voland and Beise, Chapman, Duflo, Sear 2016, parental burnout, co-residence and cohousing.
- FA-04 Raising boys and raising girls: CAH and Pasterski, Else-Quest, Todd, Voyer and Voyer, PISA 2025, Reeves, Pahlke, Dee, DuBois, Gilligan, Way.
- AN-10 Sex differences in juvenile mammals: Connellan and its meta-analytic fall, Alexander, Hassett, Pasterski, Endendijk, Eliot.
- DH-08 Ancestral child-rearing and alloparenting: Sear and Mace, Page, Boyette, van IJzendoorn, Vermeer, Thornton and McAuliffe.

CHAPTER 10

The Factor That Fell

You are sitting at a table with two people you have never met, in a university room with the blinds down, and a researcher has just put a sheet of paper between you. It shows a visual puzzle. In ten minutes it is going to be replaced by a brainstorming task, then by a moral judgement with no right answer, then by a negotiation over scarce resources in which the three of you must divide too little among too many. Nobody has told you that the point of the afternoon is not the puzzles. The point is whether the three of you, taken together, have an intelligence of your own.

That is a strange thing to ask about a table. We are used to asking it about a person. Give one person a dozen different mental tests and the scores move together: someone good at one tends to be good at the others, and the thing they have in common, which statisticians pulled out of the correlations long ago and called *g*, accounts for something like 30 to 50 per cent of the spread in any battery of tests, according to the OR-01 summary. It is one of the most replicated findings in psychology. The question Anita Williams Woolley and her colleagues asked in 2010 was whether a group has a g of its own. Give forty tables of three strangers the same battery and see whether the tables that do well on the puzzle also do well on the negotiation. If they do, a group has a single underlying ability, call it c, and you can measure it, predict with it and perhaps build for it.

They found it. Or they reported finding it, which in the years since has turned out to be a different sentence. Their c explained 43 per cent of the variance in how the groups did across the tasks, and it was only moderately

predicted by the cleverest person at the table or by the average cleverness of the three. It was predicted by how well the people at the table could read a face, by how evenly they took turns talking, and by how many of them were women. The paper ran in *Science*. The last of the three findings ran everywhere else.

I want to give you the c-factor as a tool, and then I want to show you what happened to it when two psychometricians took a second look, because the shape of what happened is the shape of everything in this chapter. A team is where the sexes meet in numbers larger than two and smaller than a parliament. It is where my mother's generation fought for the seat, and where the question of what the seat does once she has it has mostly been answered by anecdote. There is now evidence, from laboratories, from matrilineal villages and from a job advertisement sent to 7,000 people. Most of it is about the room. Some of it is about the people in the room. The two are easy to confuse, and the c-factor is where the confusion began.

One number for a room

Here is what Woolley, with Chabris, Pentland, Hashmi and Thomas Malone, actually did. Study one put 40 groups of three through the battery. Study two put 152 more groups, of three to five people, through it, 192 groups in all. The tasks were chosen from a standard map of what groups do, the McGrath task circumplex, so that no single kind of task could carry the result. The groups were scored. A single factor accounted for 43 per cent of the variance in scores across the tasks, and the average and the maximum individual intelligence in the group were only moderate predictors of it.

Three things predicted it better. The first was the group's average score on the Reading the Mind in the Eyes Test, in which you are shown a strip of photograph containing only a pair of eyes and asked which of the words offered describes what the person is feeling. The second was turn-taking: the more unevenly the talk was spread, the lower the c, a correlation of −0.41 between c and the variance in speaking turns. The third was the proportion of women. And the authors said in the paper that the third was mostly carried by the first. Women on average score higher on the eyes test,

so a table with more women had, on average, more capacity to read a face, and that was the thing that mattered.

That link needs its layer. The eyes-test advantage is one of the better-replicated small differences in the field. Kirkland and colleagues pooled 42 studies and 4,290 adults in 2013 and found it. OR-01 gives the size as g = 0.177 and, in the same sentence, as d = 0.47, which cannot both describe the same analysis, and chapter 4 settled on the smaller figure after finding only that one in the PS-02 PDF. Greenberg and colleagues found the advantage again in 2022 in *PNAS*, on 312,739 people in 57 countries, on average, at every age and across almost every language tested. OR-01 files this under physiology because it concerns the decoding of faces. I file it under behaviour, because it is a score on a test, the overlap is large, and a difference of that size leaves a great many men above the average woman at it. The point for the table is that the ability is real, it is not evenly distributed between the sexes, and it is nowhere near a rule about the man or the woman sitting next to you.

The finding spread quickly, as findings do when they tell a room what it wants to hear. Engel and colleagues replicated the eyes-test link in 2014 in *PLOS One*, in groups that worked only through text on a screen, which suggested that whatever the test measures is not just the reading of faces but something more general about reading people. A meta-analysis led by Riedl and colleagues, from the same research cluster, 22 studies, 1,356 groups, 5,279 people, reported the structure of c intact, with an average variance extracted of 43.6 per cent. By then c had been in the slide decks for years. More women, smarter teams. It was, in the Preface's terms, the second half of the question answered in one line, which should have been the first warning.

The second look

Psychometricians are the people you send for when a finding has been enjoyed too much. In 2017 Marcus Credé and Garett Howardson published a paper in the *Journal of Applied Psychology* whose title was the warning: "The structure of group task performance: a second look at the evidence for the collective intelligence construct". They took six of the published samples on which c had been built and reran the sums.

The first thing they did was calculate how much of the variance in the task scores a general factor really extracted, by the standard formula of Fornell and Larcker, and got 22 to 35 per cent, against a conventional floor of 50 for a latent factor worth the name. The second was a principal axis factoring, where a single dominant factor in data of that shape should produce an eigenvalue above about 1.38. They got 1.05, accounting for 17 per cent of the variance. The third was a meta-analysis across eight independent samples and 857 groups of how well c predicted the thing it was for, a separate criterion task the group then performed. The correlation was 0.26, with a confidence interval from 0.10 to 0.40. And they pointed out that the covariance between tasks that did exist had two cheap explanations: groups that did not try hard did badly at everything, which looks like low c and is low effort, and the data were nested in ways the original analyses had not fully handled. They cautioned, in the paper's last pages, against firms abandoning the selection of able individuals in favour of a group metric that had not yet earned it.

The factor fell. Not the finding that groups who take turns and read each other do better, which Credé and Howardson left standing and which OR-01 describes as generally accepted. What fell was c itself, the claim that a group has one intelligence the way a person has, general and measurable, and with it fell the headline that had been hung on it. Groups where a few people dominate the talk do worse. That is a finding about rooms. "More women make teams smarter" was a finding about a mediator inside a construct that did not survive its second look. Both sides are still publishing, and the Riedl meta-analysis is the proponents' reply. I give you the dispute as a dispute, which is what the title of this chapter means: the factor fell under scrutiny, and the scrutiny is not finished.

What the field studies say is messier than either side's headline, and worth having in numbers. Hoogendoorn, Oosterbeek and van Praag put about 550 business students into 45 teams of twelve in 2013 and had each team run a real company for a term, selling shares, electing officers, moving product. The sex ratio of each team was assigned. Teams near half and half outsold and outearned the male-dominated teams, and profit rose with the share of women up to about 50 per cent and then flattened, with too few heavily female teams to say what happened beyond. The authors measured conflict, friendship, decision process, atmosphere, mutual monitoring and learning, and none of it explained the result. It is the

best causal evidence that a balanced team does better at something real, and it comes with no mechanism attached.

Now the other column. Apesteguia, Azmat and Iriberri analysed a large business simulation played in teams of three by undergraduates and MBA students, in *Management Science* in 2012. Teams of three women were outperformed by every other combination. At the bottom decile the three-women undergraduate teams sat 0.47 of a standard deviation below the mean, at the top decile 0.09 below, and among the MBA teams the best results at the top came from two men and one woman. The reason was in the decisions: the all-women teams priced less aggressively, put less into research and development and more into what the simulation called social sustainability, in a game that paid for aggression. That is not a finding that women are worse at business. It is a finding that a game built to reward one end of a distribution rewards the teams drawn from that end, which is the kind of sentence that ought to be printed on the box.

And in 2024 Hardt, Mayer and Rincke put 1,368 people into 342 teams of four and had them solve a sex-neutral puzzle in an audio-only chat room, no faces, no eyes to read. All-male teams talked more, produced more words about the task and did better than both mixed and all-female teams. The striking number is inside the mixed teams. Men in a mixed room talked more than men in an all-male room, women in a mixed room talked less than women in an all-female room, and the men ended up saying almost 70 per cent more words than the women. Sort by ability and the high-ability women spoke less than the low-ability men. Take the eyes out of the room and the thing Woolley found, the even turn, goes missing. That is the boundary condition on the whole collective-intelligence story, and it is a finding about a medium, not about a sex.

Do women shy away?

Chapter 4 gave you the headline that sits under all of this: Muriel Niederle and Lise Vesterlund, 2007, 40 men and 40 women adding up numbers for three minutes, equal at the task, and 73 per cent of the men against 35 per cent of the women choosing to be paid by tournament rather than by the piece. The paper's title asked two questions, "Do women shy away from competition? Do men compete too much?", and the

data answered yes to both. Men who were bad at the task entered too often and lost money. Women who were good at it stayed out and left money on the table. The sentence to carry is the first one, because it is the one the folklore drops.

What happened to the gap under scrutiny is the pattern of this chapter again. It replicated, smaller. Dariel and colleagues pooled 53 papers in 2017 and found men entering tournaments 52 per cent of the time and women 34. Markowsky and Beblo pooled 110 studies and 409 effect sizes in 2022 and found an unadjusted gap of 13 percentage points, largest when the task was arithmetic and the subjects were students, smaller with verbal tasks, other ages and team competition. It travelled out of the laboratory. Thomas Buser, Niederle and Hessel Oosterbeek measured the competitiveness of 1,480 Dutch fifteen-year-olds in 2014 and followed them into the four school tracks the Dutch system offers, and the laboratory measure accounted for up to 23 per cent of the sex gap in who chose the prestigious, mathematical tracks, with ability, risk and confidence held constant. Buser found in 2016 that after a loss, women cut their effort whatever the prize and men only when the prize was large, which explained about 11 per cent of the variation in women's effort and 4 per cent of men's, and roughly half of the sex gap in repeated competition. That is the behavioural layer, and it is statistical. By Julie Nelson's calculation the two distributions overlap by 80 to 90 per cent, and a random woman is the more competitive of a random pair 40 to 45 times in a hundred.

Then came the measurement people. Ben Gillen, Erik Snowberg and Leeat Yariv argued in *Econometrica* in 2019 that the preference for competition Niederle and Vesterlund had found after controlling for risk attitude and confidence was an artefact of measuring risk attitude and confidence badly. Measurement error accounts for 30 to 40 per cent of the variance in those elicited traits, they estimated, and a control measured with error shrinks toward zero and leaves its residue to be picked up by whatever is left in the regression, which was the variable for sex. When they reran the design on the Caltech Cohort Study with duplicate elicitations as instruments, the sex gap in entry was entirely accounted for by risk attitude and overconfidence. There was no separate shyness. There were women who were on average more risk-averse and more accurately confident than men, and that was enough. Price ran a pure replication in 2020 and got nothing at all. Whether the gap is a preference or a symptom is, OR-03 says, open.

It is a different kind of open from the c-factor's. Nobody disputes that fewer women enter. They dispute what the entry is made of.

The hormone that was supposed to be underneath it has been looked for and not found, and chapter 3 owns that story, so one paragraph. The pre-registered trial that settles the question for now is Bird and colleagues, *PNAS*, 2024: 1,000 men, 11 milligrams of testosterone up the nose or a placebo, nine hypotheses about competition, risk and social preference, power to detect an effect of 0.23 of a standard deviation, and no effect on anything. Zethraeus and colleagues had found the same in 200 postmenopausal women in 2009. Dekkers and colleagues pooled 52 studies and 94 measurements in 2022 and found the correlation between testosterone and risk-taking sitting on zero, with a confidence interval from −0.043 to 0.039. The subtler idea, that testosterone drives status-seeking only when cortisol is low, the dual-hormone hypothesis, got its meta-analysis too: 30 papers, 33 studies, 8,538 people, an interaction of r = −0.061, significant and negligible. OR-02 attributes that meta-analysis to Dekkers and colleagues, 2019, in *Hormones and Behavior*, and PS-05 to Geniole and colleagues, 2016, in *Psychoneuroendocrinology*, with identical counts, and PS-05's own citation links to a paper carrying the Dekkers title, so I carry Dekkers and mark it. Chapter 11 returns to the hypothesis at the scale of the boardroom. For the table, the point is this: whatever makes 73 per cent of the men enter, a dose of the molecule does not move it.

The Khasi and the form

Culture, the third layer, is where the gap does the thing averages are not supposed to do. It changes sign. Uri Gneezy, Kenneth Leonard and John List took the tournament choice to two societies in 2009, in *Econometrica*: the Maasai of Tanzania, where women hold little property, and the Khasi of Meghalaya in north-east India, where inheritance and clan membership run through the mother and the youngest daughter takes the family wealth. Chapter 4 gave the figures: half the Maasai men compete and a quarter of the women, 54 per cent of the Khasi women and 39 per cent of the men. The task was tossing a ball into a bucket. The Khasi women competed more than the Maasai men. Binglin Gong and Chun-Lei Yang found the same reversal in generosity and risk among the

matrilineal Mosuo and patrilineal Yi of Yunnan, and then watched it fade: Mosuo girls started out more willing to take risks than Mosuo boys and, after years in schools shared with the Han majority, were more risk-averse than the boys by the age of eleven. Culture did not just set the gap. It reset it inside a childhood. And in 2023 Karen Hauge and Andreas Kotsadam measured willingness to compete in 1,943 Norwegians whose parents came from 59 countries, people raised under the same schools and the same labour laws, and found the gap smaller the more equal the parents' country of origin, from both sides: the men from equal countries competed less, the women more.

So the gap is real, replicable, smaller than the first paper, possibly made of risk and confidence rather than of shyness, moved by no hormone that has been tried, and reversible by kinship. What can a team do with that? The answer the economists found is almost insultingly practical. Change the form.

Flory, Leibbrandt and List placed a real job advertisement in front of about 7,000 job seekers in 2015, in *Management Science*, and varied one thing. In one version the pay was competitive if you ticked the box to make it so. In the other the pay was competitive unless you ticked the box to make it not. With the competition as the thing you had to choose, 72.5 per cent of the men applied and 46.7 per cent of the women, a gap of nearly 26 points, about the size of Niederle's. With the competition as the default you could leave, 76.4 per cent of the men applied and 75.3 per cent of the women. The gap did not shrink. It went. Nobody's preference changed. The form did what a generation of advice to women had not, by not asking. Niederle, Segal and Vesterlund found in 2013 that a quota guaranteeing women half the winning places pulled in capable women who had been staying out, and did not lower the standard a winner had to meet, because the women it drew in were good.

The rest of the architecture of a workplace carries the same lesson. Hannah Riley Bowles, Linda Babcock and Lei Lai showed evaluators in 2007 the files and videos of candidates who had either accepted a salary or asked for more, and the evaluators penalised the women who asked far more than the men who asked, a penalty OR-03 says was nearly three times the size, with the figures stripped from the text and the PDF alike. When the subjects were put in the candidate's chair, the women declined to negotiate with a male evaluator, and what explained it was the anticip-

ated cost, not a lack of nerve. Christine Exley and Judd Kessler had more than 4,000 adults and 10,000 schoolchildren do a maths and science task in 2022 and then rate their own performance from 0 to 100. With actual performance held constant, women rated themselves 13 points lower than men, about a quarter of the average rating, when the rating was private as well as when it was not, and already in the sixth grade. Employers who saw the self-ratings hired the women less and paid them less. And Linda Babcock, Maria Recalde, Lise Vesterlund and Laurie Weingart found in the *American Economic Review* in 2017 that women volunteered 50 per cent more often than men for the task that helped the group and cost the volunteer, about 20 percentage points, that managers of both sexes asked women to do it more often, and that women said yes more often when asked. Everyone in the room expected the woman to do the unrewarded thing, including the woman.

Every one of those is a finding about a form, a default, a chair or an expectation, with a behavioural difference underneath it that is real on average and overlapping. The three-layer sentence for a team is this: the capacity to read a face and the appetite for a tournament are measurable and differ by sex at the level of the population. What they do at the table is statistical. What the table asks of them is ours.

The room decides

Scale it up once, to the bodies that decide for others, and the same shape holds. Chattopadhyay and Duflo used a 1993 Indian constitutional amendment that reserved a third of village council chief posts for women, drawn by lot, and compared 265 councils in West Bengal and Rajasthan in 2004, in *Econometrica*. Councils led by women put more into drinking water, and in West Bengal into roads, less into schooling, and women in those villages were twice as likely to speak in a council meeting. That is as close to a randomised trial of a sex in power as politics has produced, and chapter 14 owns the state, so I leave it there.

The boardroom is more instructive for a team, because its headline fell in the same way the c-factor did. Norway required 40 per cent women on company boards from 2003. Ahern and Dittmar reported in 2012, in the *Quarterly Journal of Economics*, that the quota had cut firm value by

12.4 per cent, a figure that went round the world. Eckbo, Nygaard and Thorburn reran the event study in 2022 and found the original had dated the events wrongly and had not corrected for the fact that every firm's shares moved together on the day the law hit them all. Corrected, the effect was indistinguishable from zero. Post and Byron's meta-analysis of 140 studies in 2015 found women on boards associated with slightly better accounting returns and with nothing much on the share price. Sila, Gonzalez and Hagendorff found no effect on equity risk at all in 2016. And in the room where risk is the whole job, the monetary policy committee, Masciandaro, Profeta and Romelli found in 2020 that a higher share of women went with higher interest rates for the same expected inflation. The women were the hawks, which was not in anybody's pamphlet. The authors' guess is selection, the only women who survive the profession's filters being the hard ones, or signalling, the need to prove you are not soft. Either way the baseline average was not in the room, because the room had chosen who came in.

One more study, which is not about people. Couzin and colleagues trained a minority of golden shiners to swim toward a yellow target and a majority toward a blue one, in *Science* in 2011. Minority and majority alone, the minority steered the school, because it wanted it more. Add five to ten fish with no training and no preference, and the school went where the majority wanted. The uninformed restored the majority. Conradt and Roper had shown in *Nature* in 2003 why a group might evolve to share its decisions: shared decisions are less extreme, and the cost of following an extreme one falls on everyone. A fish tank is not a boardroom, though some boardrooms have been compared to one unkindly. The lesson survives the move. A body's decisions are shaped by who is in it and by how many in it hold no brief, and the second is a design choice.

Here is the turn. Everything in this chapter that fell, fell because somebody had read a composition as a cause: more women, more c. More women on the board, less value. More testosterone, more entry. Everything that held was about the room: the turn, the medium, the default on the form, the chair you ask from, the task nobody is paid for, the fish with no opinion. The accent is in the people, and it is real. The factor is in the room, and the room is the thing you can change by Monday.

THE STRONGEST CASE AGAINST

The chapter has argued that the sex of the people at the table matters less than the rules of the table. Here is the case that it matters more than I have let it, in its own numbers.

Markowsky and Beblo's 110 studies still show a 13-point gap in tournament entry after every design is averaged in, and Dariel's 53 papers show 52 against 34. The Gillen critique explains the gap, it does not remove it: women are more risk-averse and more accurately confident on average, and whichever word you put on that, fewer of them enter. Buser, Niederle and Oosterbeek's 1,480 Dutch fifteen-year-olds carried the laboratory measure into their choice of school track, at 23 per cent of the sex gap, before any employer had seen them. Exley and Kessler's 13-point gap in self-rating was there when nobody was watching and was there in eleven-year-olds. Buser's 2016 result on losing, 11 per cent of the variation in women's effort against 4 in men's, is a difference in how the sexes carry a defeat into the next round, and no form fixes the next round. Byrnes, Miller and Schafer's 150 studies put men ahead on risk in 14 of 16 domains, and chapter 4 gave the financial figure at about 0.43 of a standard deviation.

Then the teams themselves. Apesteguia's all-women triads scored below every other combination at a business game, by nearly half a standard deviation at the bottom of the distribution, because of the decisions they took. Hardt's all-male teams beat mixed and all-female teams on a sex-neutral puzzle. Hoogendoorn's balanced teams did better, but the profit curve flattened above 50 per cent women and the authors could find no mechanism in anything they measured. Saccardo, Pietrasz and Gneezy's finding that the competition gap vanishes among people in science and engineering, which chapter 4 carried on 97,678 adults, is on its authors' own reading a story of selection: the competition-averse women never arrive. The Khasi reversal is one society, the Mosuo reversal was erased by a decade of school, and the Norwegian result runs in the same direction as the paradox chapter 5 left open. Flory's opt-out default equalised who applied. It said nothing about who performed or who stayed. And Woolley's three correlates were never refuted as behaviour: Credé and Howardson left standing the finding that groups with higher average eyes-test scores and more even turns do better, and the eyes-test advantage now rests on 312,739 people in 57 countries. On that reading the women at the table were carrying something real, the c-factor was only a bad name for it, and a chapter that files the whole thing under "the room" has mistaken the

container for the contents.

There is a second jaw, from the other shelf, and it bites the same chapter from the other side. SY-01 reports a meta-analysis by Mazei and colleagues of 10,888 negotiators in which the male advantage in negotiation vanishes when people negotiate on behalf of others rather than for themselves. If that holds, then even the "real on average" difference the first jaw rests on is a difference in who is allowed to ask for what, and the room is all there is.

I think the first reading is about half right, and I have given it the numbers so you can decide which half. What neither reading can do yet is name the mechanism. Nobody has found the mediator in Hoogendoorn's data, the hormone in Bird's, or the preference in Gillen's. Until someone does, the room is the only part of this a team can act on.

IN PRACTICE

If your team is small, three or four, and one of you is the only woman or the only man, then the research on tokens applies: solo women in majority-male groups talk less and carry less weight (Bear and Woolley 2011, per OR-01), and in Hardt's audio rooms the gap in words reached 70 per cent. The practice is to put the turn-taking in the structure, not in the goodwill: a round in which each person speaks before anyone speaks twice, a written first pass before the spoken one. The failure mode is the round that becomes a ritual while the real decision happens afterwards in the corridor, which is where it was happening before.

If your team works by voice alone, by phone or in a chat with the cameras off, then expect the Hardt pattern, men speaking more than they do in a room of men and women less than they do in a room of women, and expect the able women to be the quietest. Turn the cameras on, or move the first round into text, where Engel's groups still read each other. The failure mode is that Engel's result is one study, and the counter-evidence is Hardt's: the medium may decide more than any rule you add to it.

If you choose who gets the hard assignment or the promotion by asking people to put themselves forward, then you are running Niederle's tournament with the opt-in form, and in Flory's field trial 46.7 per cent of the women applied under that form against 75.3 under the other, before anyone was assessed. Make the competitive track the default and let

people leave it. In Flory's data the gap closed to a point. The failure mode is that opt-out equalises entry, not outcome: on Gillen's reading the women who stayed out were right about their own risk, and some who are now in tend to lose, as the low-ability men in Niederle's data did.

If pay in your team is set by negotiation, then the Bowles penalty falls on whoever is seen as asking against type, and in 2007 that was women, by about three to one. Publish the bands. The failure mode is that a band removes the ask and with it the one lever the strongest negotiator had, and nobody has measured what a team loses when its best bargainers cannot bargain.

If the tasks that keep the team running but promote no one are filled by asking who is willing, then both sexes expect a woman to say yes, the manager asks her more often, and she says yes more often (Babcock and colleagues 2017, about 20 percentage points). Rotate the task by list or by lot. The failure mode is the list that quietly exempts the person whose time is deemed too valuable, which restores the old allocation under a new name, and the counter-evidence is that nobody has shown rotation improves a team's output, only its fairness.

If you read self-assessments, then trust the task and not the rating: the 13-point gap in Exley and Kessler's data was in the private ratings too, and the employers who used the ratings hired wrongly. Where you must use the rating, tell whoever reads it what the gap is. Telling advisers closed a 12-point gap in the ambition of the advice women got. The failure mode is the reader who, told "women underrate themselves", adds 13 points to every woman, which is statistical discrimination in the other direction and wrong for a large share of them, if the distributions overlap as Nelson says they do.

If the team is deciding something with a real downside, money, a merger, a move, then there is one honest reason to want it mixed, and it is not that women are careful. It is the fish. A body whose members do not all want the same thing hard is harder to steer off a cliff, and a body in which some members hold no brief steadies toward its majority (Couzin 2011). The failure mode is the central bank. Masciandaro's women were the hawks because of who gets through the door, so if your door selects for one temperament, mixing the sexes behind it tends to change nothing you can measure, and Sila's boards showed no change in risk at all.

If you were hoping to make the team smarter by adding women, then

do not promise it to anyone. The c-factor that underwrote the promise explained 17 to 35 per cent of the variance on the second look and predicted the next task at r = 0.26. What you can say is that a team in which everyone speaks, and in which people can read each other, tends to do better across tasks, and that the share of women in a room has, on average, moved both of those in the laboratory. On average. The man across the table may be the one who reads the room best. Look at him, not at the ratio.

And if you are the one woman at the table, or the one man, then the evidence says the room is reading you as a category before it has heard you, and that the quietest version of you is the one it is likely to get. The one thing that reliably changed that in the data was a rule that made the turn yours before you had to take it. Ask for the rule. The failure mode is the room that gives you the turn and takes the decision elsewhere, and the counter-evidence is Hoogendoorn's: in teams large enough that nobody was the only one, no rule was needed, because the numbers did the work.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. Woolley 2010's three correlates of group performance as behaviour (OR-01, PS-05). Credé and Howardson 2017's second look: 22 to 35 per cent variance extracted, eigenvalue 1.05, r = 0.26 on 857 groups (OR-01). Hardt 2024's audio-room asymmetry (OR-01). The tournament-entry gap at meta-analytic size, Dariel 2017 and Markowsky and Beblo 2022, and its external validity in Buser 2014 (OR-03, PS-05). Gillen 2019's measurement-error account (OR-03). Bird 2024 and Dekkers 2022 nulls (OR-03). Gneezy 2009's reversal and Hauge and Kotsadam 2023's dose-response (OR-03). Flory 2015's opt-out default (OR-03). Eckbo 2022 overturning Ahern and Dittmar 2012 (OR-01, SO-07).

Decoration. The blinds-down room in the opening. The ball and bucket. The box the game should be printed on. The golden shiners. The hawks at the central bank. Chattopadhyay and Duflo's councils, which belong to chapter 14.

Contested. Whether c exists at all (Woolley, Riedl against Credé and Howardson). Whether the entry gap is a preference or risk plus confidence measured with error (Niederle and Vesterlund against Gillen, Snowberg and Yariv, with Price 2020's null). The Kirkland effect size, 0.177 or 0.47. The dual-hormone meta-analysis attribution, Dekkers 2019 or Geniole 2016. Whether Hoogendoorn's premium has a mechanism. Whether Saccardo's vanishing gap is selection or context. Whether the Eckbo correction is the last word on Norway. Whether the Mazei advocacy result (not in this chapter's documents) removes the negotiation difference entirely.

NUMBERS TO CARRY

- Woolley's c explained 43 per cent of variance across tasks and correlated at −0.41 with the variance in speaking turns, in 192 groups (Woolley et al. 2010, *Science*, per OR-01). On the second look, 22 to 35 per cent extracted, an eigenvalue of 1.05, and r = 0.26 (CI 0.10 to 0.40) with criterion tasks across 857 groups (Credé and Howardson 2017, *Journal of Applied Psychology*, per OR-01).
- Men said almost 70 per cent more words than women in mixed four-person audio teams, 1,368 people, 342 teams (Hardt, Mayer and Rincke 2024, *Management Science*, per OR-01).
- Tournament entry: 73 against 35 per cent in the original 80 people (Niederle and Vesterlund 2007, *QJE*), 52 against 34 across 53 papers (Dariel et al. 2017), and a 13-point unadjusted gap across 110 studies and 409 effect sizes (Markowsky and Beblo 2022), all per OR-03.
- Opt-in: 72.5 per cent of men and 46.7 per cent of women applied. Opt-out: 76.4 and 75.3. About 7,000 job seekers (Flory, Leibbrandt and List 2015, *Management Science*, per OR-03).
- Testosterone and competition: 1,000 men, 11 mg intranasal, nine hypotheses, no effect (Bird et al. 2024, *PNAS*). Testosterone and risk, 52 studies, 94 measurements, CI −0.043 to 0.039 (Dekkers et al. 2022, *Neuroscience & Biobehavioral Reviews*), both per OR-03.

THE LIBRARY

- OR-01 Mixed-sex teams and collective intelligence: Woolley, Credé and Howardson, Riedl, Engel, Hoogendoorn, Apesteguia, Hardt, Kirkland, Greenberg, Bear and Woolley, Lau and Murnighan, Adams and Ferreira, Ahern and Dittmar, Eckbo, Post and Byron.
- OR-03 Competition, negotiation and promotion systems: Niederle and Vesterlund, Dariel, Markowsky and Beblo, Buser, Gillen, Price, Bird, Zethraeus, Dekkers 2022, Knight, Gneezy, Gong and Yang, Hauge and Kotsadam, Flory, Niederle Segal and Vesterlund, Bowles, Exley and Kessler, Babcock 2017, Nelson.
- PS-05 Competition, status, risk and hierarchy: the dual-hormone meta-analysis as Geniole 2016, Byrnes, Buser 2016, Saccardo, Woolley and Apesteguia restated, Hyde 2005.
- SO-07 Mixed vs single-sex decision bodies and collective risk: Eisenegger, Couzin, Conradt and Roper, Chattopadhyay and Duflo, Ahern and Dittmar, Matsa and Miller, Eckbo, Sila, Post and Byron, Masciandaro, the Capital Jury Project.
- OR-02 Leadership, hormones and sex (consulted for the dual-hormone attribution only): Dekkers 2019.

PART THREE

Emancipation 2.0

CHAPTER II

The Abbess and the Monks

A woman of fifty-one is chairing a meeting on the fourth floor of an office building and her face has started to burn. It began at the collarbone a moment ago and is climbing. The room is at 21 degrees and everyone else in it is comfortable. She does not touch her collar or reach for the water, because she has learned that the one thing worse than the heat is being seen to notice it. Across the table a man of forty-four is in the fifty-sixth hour of his working week, the third week this quarter over fifty-five, and he does not notice that either. He is proud of it, in the quiet way men of my generation are proud of things we were taught not to mention.

Neither body is doing anything unusual. Hers is finishing its reproductive years at the age the British average does it, 51 (OR-04). His is running a working week that, in the largest pooled cohorts in existence, goes with a third more strokes than a forty-hour one (Kivimäki and colleagues, *The Lancet*, 2015, 528,908 people). The room was built for neither of them. It was built for a body that is never cyclical, never pregnant, never tired, and has somebody at home. That body does not exist. It has run most organisations for about two hundred years.

Think of the last organisation you belonged to that had more than a dozen people in it, and ask it the question this chapter asks of every organisation in the record. Whose body was the rule written for?

Nine centuries ago a wandering preacher wrote a rule for a house in which the men took their orders from a woman. Robert of Arbrissel founded the abbey of Fontevraud, in France, in 1101. It had a house of monks and a house of nuns, kept strictly apart, and both answered to the abbess.

The monks owed her, the rule said, the obedience a son owes his mother. The arrangement outlived the Hundred Years' War and the Wars of Religion and was still standing when the French Revolution closed it in 1790 (OR-06). Six hundred and eighty-nine years is a long time for an organisation to hold a difference between its members and an asymmetry of authority at once.

That house is the chapter's picture. Not because a woman in charge is the point, which it is not. Because its rule answered the two questions every organisation has to answer, who does what and who answers to whom, as two separate questions. Most of the organisations in this chapter answered them as one.

The hormone that was supposed to run the house

Start where the organisation chart was supposed to be written, in the measurable layer.

In 2010 Mehta and Josephs published, in *Hormones and Behavior*, what became known as the dual-hormone hypothesis. Testosterone alone had never made the boss, so the new idea was a pair: testosterone drives a person toward status, but only when cortisol, the stress hormone, is low, which has an attractive logic to it, since the body that is busy surviving should not be busy climbing. The report that still believes this tells the flagship study, corporate executives whose testosterone predicted the number of people they managed, but only among the low-cortisol ones (OR-05, which does not name the study).

Chapter 10 carried the meta-analysis that took it apart and the puzzle about who wrote it, so here is one line. Thirty papers, 33 studies, 49 effect sizes, 8,538 people, and an interaction of $r = -0.061$, in the predicted direction and too small to use for anything (Dekkers and colleagues, *Hormones and Behavior*, 2019, per OR-02). What this chapter adds is what came after. Grebe and colleagues ran a p-curve over the published literature in the same year and found it had about 16 per cent statistical power to detect the effects it claimed. Then they opened the file drawer, seven unpublished datasets, and found no interaction in any of them (OR-02).

There is a second thing wrong with the hypothesis, which matters more for this book. It was built on men. The standard immunoassays

are unreliable at the low testosterone concentrations women run at, so a weaker effect in women may be a weaker measurement (OR-02). And oestradiol, which might do for a woman's drive to status what testosterone was supposed to do for a man's, is barely in the literature at all. The one hormone that was studied failed to run the boardroom. The other was not invited.

I owe you the dissent. One of the six reports under this chapter treats the dual-hormone hypothesis as live and goes on to a 2026 study by Huang and colleagues of 826 working men aged 22 to 40, in which the high-stress group had a 54.37 per cent prevalence of erectile dysfunction, ten times the odds of the low-stress group, with cortisol rising first, testosterone falling two to four weeks later and function four to six weeks after that (OR-05, which cites a press release). I carry both reports and I do not think they are about the same thing. The dual-hormone hypothesis says the hormones pick the leader. Huang says the job picks the hormones. The second survived. The first did not.

So the measurable layer, at this scale, measures a body answering to its organisation. Which is the Endocrine Accent from the other side: the lilt is real, and the organisation is one of the things that sets it.

What the abbess does differently

Now the behaviour, which is statistical, and here the two bells are very nearly one bell.

Alice Eagly and Johnson pooled 162 studies of leadership style in *Psychological Bulletin* in 1990 to test the stereotype everyone knew, that women lead by relationships and men by task. In real organisations the task-oriented difference came to $d = 0.00$ and the interpersonal one to 0.04. The one difference that survived was style of decision: women leaned democratic and participative, men autocratic and directive, at $d = 0.22$, which on chapter 4's table is about 90 per cent of the two distributions shared (OR-02). Vecchio objected in 2002 that 197 of the primary studies were leaders rating themselves, which is to say leaders describing the leader they thought they were. Fair.

Does any of it show in results? Paustian-Underdahl, Walker and Woehr put 99 samples and 101,676 people into one analysis in the *Journal of Ap-*

plied Psychology in 2014 and found no difference in rated effectiveness, $d = -0.05$. Then they split the raters. Men rated themselves better leaders than women rated themselves. Everyone else, bosses, peers and the people underneath, rated the women better, at $d = -0.12$ (OR-02). The authors' reading is a filter: a woman has to clear a higher bar to be given the job, so the women who hold it are on average the stronger ones. Whatever the explanation, a difference that appears when you ask the leader and reverses when you ask the led is not a difference in leading.

Where the sexes do differ is in who gets to lead in the first place, and that difference lives in the room. Badura and colleagues took 1,632 effect sizes from 120 samples, 32,579 people, in the same journal in 2018 and asked who emerges as leader in a group that starts without one. Men, still, though the gap has been closing for decades. The mechanism they propose is talk: assertive and dominant people speak more, the group hands the chair to whoever has been speaking, and those traits run higher in men on average (OR-02). The counter-mechanism is in the same data. Once a woman has the chair, the ability to read the room, which runs higher in women on average, predicts that the group rates her leadership as the good kind. Emergence rewards the accent of agency. Performance rewards the other one. An organisation that promotes by emergence has chosen, without noticing, which it is paying for.

Then the cultural layer, which Eagly also named. Role congruity, from Eagly and Karau in *Psychological Review* in 2002, is the finding that the leader stereotype and the female stereotype point different ways, so a woman who leads the communal way is a weak leader and a woman who leads the agentic way is an unpleasant woman (OR-02). The same scripted autocratic behaviour was rated worse in a woman than in a man across Eagly, Makhijani and Klonsky's meta-analysis of 1992. Koenig and colleagues pooled 40, 22 and 17 studies across three paradigms in 2011 and found the leader stereotype still masculine, weakening slowly, and weakest in education, which is where the women already were. And Ryan and Haslam, reading FTSE 100 board appointments in the *British Journal of Management* in 2005, found women appointed to the top disproportionately when the firm was already failing, the glass cliff, which Haslam and Ryan in 2008 traced to a heuristic, think crisis, think female, and to a scapegoat's availability. The cliff is the only chair on offer.

That is the behavioural layer. Nearly nothing in the leading, something

in the emerging, a great deal in the watching.

The rule of the house

Now the history, which is where this chapter has its only real advantage over the two shelves, because the shelves argue about what an organisation would look like if it were built around the difference, and the record has several.

The Anglo-Saxon church built one first. Between about 600 and 900, Sarah Foot shows in *Monastic Life in Anglo-Saxon England* (2006), the double monastery, monks and nuns in one foundation under one abbess, was the mission station of the English conversion, and the abbess was very often a royal widow or a king's daughter, which fused the church's authority with the crown's (OR-06). Hild of Whitby is the name that survives. The theological cover was that people who had given up sex, property and will had given up their sex as well. Jo Ann McNamara's objection is that the chroniclers praised the great ascetic women for acting *viriliter*, like a man, so what the cloister offered a woman was not neutrality but promotion to honorary man (OR-06). I think she is right and I think it was still the best deal available in 700.

The East said no. The Second Council of Nicaea in 787, in its twentieth canon, ruled that women ought not to live in bishops' houses nor in monasteries of men, and banned new double houses in the Orthodox world (OR-06). The West did not follow, and in 1101 Robert of Arbrissel, an itinerant preacher, built Fontevraud on the model the East had just forbidden and then improved it. Separate houses, strictly kept. One abbess over both. Obedience of the monks to her modelled on a son's to his mother, with the Virgin as the warrant (OR-06). The Gilbertines at Sempringham in England went further on the small things, where rules live: exactly the same food and drink for the nuns as for the canons (OR-06). Separate quarters. One table. One rule. And the rule said who answered to whom in so many words, so that nobody had to find out by seeing who spoke most.

The guilds are the same question answered the other way. Sheilagh Ogilvie's *The European Guilds* (2019) sits on more than 5,000 quantitative observations and 12,000 qualitative ones across eight centuries, and its

finding is that guilds were cartels that colluded with the political class to restrict competition, and that one of their standard methods of keeping members' wages up was to keep women out (OR-06). Her earlier work on Württemberg makes the point that matters for the layers: the guilds had to pass regulation after regulation against women weaving, brewing and trading, and a rule that has to be re-passed every generation is a rule the bodies keep breaking (HC-04). Judith Bennett's court rolls from Brigstock in Northamptonshire show brewing in the early fourteenth century run by more than 300 women, a third of the adult women of the village. When hops arrived and beer kept and the trade needed capital, it became a learned art, and learned arts were for men (Bennett 1996, per OR-06). The exception proves the mechanism. In 1675, with Colbert behind them, the seamstresses of Paris got a guild of their own, with a monopoly on clothing for women and for children, boys up to eight. By the 1750s it was taking 419 apprentices a year under more than 2,000 mistresses, the largest apprentice body in the city of either sex (Clare Crowston, *Fabricating Women*, 2001, per OR-06). Given the licence, the women ran a cartel exactly as well as the men. Which is not a compliment to either.

Below both, most people lived in a third arrangement with no rule written at all, the working couple of Heide Wunder's early modern Germany, his portfolio the plough and the burst strength, hers the dairy, the textiles and the market, and neither solvent without the other (*He Is the Sun, She Is the Moon*, 1998, per OR-06). Three designs, then, before a single factory. Separate houses under one rule that named the authority. A cartel that wrote the body into the rule. And a couple that wrote nothing and divided by what each body did best. The chapter is about which of them we rebuilt.

The worker with no body

We rebuilt the cartel, and gave it a factory.

The pre-industrial household was the firm, with subsistence, market production and children in one place, and industrialisation moved the first two out of the house and left the third behind (HC-04). Claudia Goldin's U-shaped curve, drawn across 162 countries and the American census from 1890 on, says what that did to women's paid work: high in

farming economies, falling through early industrialisation as work left the home and a rising male wage bought a wife's absence from the labour market, rising again only when offices wanted education more than muscle (Goldin 1995, *World Development*, per HC-04). Leonore Davidoff and Catherine Hall read the wills, sermons and house plans of the English middle class in Birmingham and Essex between 1780 and 1850 and found the ideology poured to fit the shape: commerce as male and corrupting, the home as female and the nursery of virtue, a wife's idleness as the mark of a man who had arrived (*Family Fortunes*, 1987, per HC-04). Whether the working class wanted the same thing is the oldest fight in the field, Heidi Hartmann in 1976 reading the family wage as a pact between capital and patriarchy and Jane Humphries in 1977 as one room the factory could not reach (HC-04). In 1,350 British household budgets from 1787 to 1865 the breadwinner mostly failed to appear (Horrell and Humphries, per HC-04).

Out of that came the body the rule was written for. Jane Lewis's 1992 typology of welfare states sorts countries by how hard they built the breadwinner into tax and childcare, and Britain and Germany are her strong cases, Sweden and Denmark her weak ones (HC-04). I was born in 1975 into the strong column. Chapter 13 has the tax code. What I want here is the worker the strong column assumed: paid, continuous, available, with the laundry done by someone else, whose work Ivan Illich called shadow work (HC-04). Sociologists call him the ideal worker. Here is what he costs the two bodies that actually turn up.

Her first. Schoep and colleagues asked 32,748 Dutch women about their periods in *BMJ Open* in 2019. Thirteen point eight per cent had stayed home for them, on average 1.3 days a year. Eighty point seven per cent had gone to work in pain and reported a third less output on those days, which over a year came to 8.9 days of work lost without a single absence on the record (OR-04). The absence the rule can see is small. The presence it cannot see is seven times larger, self-reported, and not yet measured from the employer's side (OR-04, its own caveat). Now the laws. Japan has had menstrual leave since 1947, and a 2020 ministry survey found 0.9 per cent of eligible women using it. South Korea mandated a paid day a month in 2001 and a 2018 poll found about 19 per cent taking it. Spain in 2023 began paying for it from the state, at 60 per cent of base pay, on a medical diagnosis (OR-04). Jonathan Gruber showed in the *American*

Economic Review in 1994 that when a state mandates a benefit that only one sex uses, employers price it in and the cost comes back to that sex as lower wages and fewer offers, and Buriakova, Fitzpatrick and Machin found the same across the OECD in 2022. The leave is a cost on her payslip before she has taken a day of it.

The night shift is the clearest case in the chapter of a design decision landing on a body. Wang and colleagues pooled 21 studies and 195,538 women in 2019: shift work went with 1.30 times the odds of irregular cycles, 1.35 times the odds of painful ones and a small rise in the hazard of early menopause, 1.09. Bonde and colleagues reviewed 30 papers in the *Scandinavian Journal of Work, Environment & Health* in 2013 and found fixed night shifts going with 1.51 times the risk of miscarriage across the five rigorous studies. In a 2019 review, more than 55.5 hours a week went with 1.21 times the odds of a preterm birth (OR-04, which calls this a 10 per cent increase in the same sentence). And Wallace and colleagues, in a cohort of 5,191 women in 2023, found about 75 per cent higher odds of gestational diabetes in afternoon and night workers (OR-04). Nobody ordered any of this. Someone wrote a rota.

Then the end of the cycle, which is where the woman at the table started this chapter. About 90 per cent of women have symptoms through the transition. The Fawcett Society asked 4,014 British women in 2022: 10 per cent had left a job because of them, 14 per cent had cut their hours, 8 per cent had turned down a promotion, 44 per cent said their ability to do the job had been compromised, and 41 per cent said colleagues had treated it as a joke (OR-04). The British Standards Institution published a standard for it in May 2023, BS 30416, which is unglamorous and specific: breathable fabrics, less time in heavy protective kit, a fan, a window, shifts moved on request (OR-04). Against which Grandey, Jack and Brewis warn that a firm that files menopause under medicine files the women under decline, an older man experienced and an older woman symptomatic (OR-04). Both are right. The window and the register are different things.

His body now, and his is the one this book's shelves usually leave out. Mika Kivimäki and the IPD-Work consortium pooled 25 studies from Europe, the United States and Australia in *The Lancet* in 2015: 603,838 people followed for coronary heart disease for 8.5 years, 528,908 for stroke for 7.2. Against a 35 to 40 hour week, 55 hours or more went with a relat-

ive risk of 1.33 for stroke, confidence interval 1.11 to 1.61, and 1.13 for heart disease, climbing in steps from 41 hours up, after adjustment for smoking, drink, inactivity and class (OR-05). The World Health Organization and the International Labour Organization then did the sums. Pega and colleagues, in *Environment International* in 2021, from more than 2,300 surveys in 154 countries, put the deaths from heart disease and stroke attributable to working 55 hours or more in 2016 at 745,194. Seventy-two per cent of them were men. The two agencies named long hours the occupational risk factor with the largest attributable burden of disease (OR-05). Not asbestos. Not the mine. The week.

The mine is in there too. In the United States there were 5,283 fatal work injuries in 2023, and in every year from 2011 to 2024 men were between 91.4 and 93.0 per cent of the dead, from about 53 per cent of the employed, a fatality rate more than nine times women's that has not moved in a generation of safety law (OR-05, from the Bureau of Labor Statistics). The reason is sorting, not daring: structural iron, roofing, fishing, logging, construction and the heavy roads are where the men are.

And when the work stops. Milner, Page and LaMontagne pooled the cohort studies in 2013 and found unemployment going with 1.70 times the risk of suicide, 1.15 once prior mental health was allowed for, 2.50 in the first five years, and higher in men than women throughout (OR-05). Chapters 6 and 7 have what a husband's lost job does to a marriage, so I only point at it. The provider role is a cultural script. It is written on a body that reads it as a verdict.

The script runs the other way too. A survey of nearly 1,000 fathers in large American firms found 95 per cent calling flexibility critical and 53 per cent saying they would consider staying home full-time if the money allowed (Boston College Center for Work & Family, *The New Dad*, per OR-05). Rudman and Mescher showed in 2013 why they do not: a man who asks for family leave is read as less committed and, specifically, less masculine, and is rated and promoted accordingly, the flexibility stigma (OR-05). Men given flexibility often work more hours than before, to prove they did not want it. So the ideal worker is kept in place from both sides. She is priced for a body she has. He is penalised for asking to have one.

Separate doors

The last question is the oldest: when the bodies differ this much, do you build one house or two?

The answer in the record depends entirely on which layer the task lives in, and the military is the cleanest test because it runs on all three. Robert MacCoun's review for RAND in 1996 took apart the word cohesion, the whole argument against mixing: across 66 correlations in 49 studies, what predicts a unit's performance is task cohesion, the shared commitment to the job, not social cohesion, whether the members like each other (OR-07). A Norwegian study by Hellum in 2014 found that putting conscripts of both sexes in the same dormitory took the drama out of sex and built task bonds (OR-07, cited to a secondary source).

Then the task crossed into the measurable layer. In 2015 the United States Marine Corps ran a year-long experiment, the Ground Combat Element Integrated Task Force, with 100 female volunteers in otherwise closed combat units. Of 134 tasks observed, 93 showed a statistically significant difference between the all-male and the mixed teams, and in 88 of the 93 the all-male teams did better. Thirty tasks showed a gap of 30 per cent or more, 28 of them in the men's favour, and they were the ones that matter most: carrying a casualty, marching under load, clearing an obstacle. Women were injured more. At the Infantry Training Battalion, 36 to 46 per cent of women graduated, against about 98 per cent of men (OR-07). Israel's mixed battalion, Caracal, gives the other half: among 85 men and 235 women followed for three years, the women had more stress fractures and lower attrition, 28 per cent against the men's 37, because the women who chose it had chosen it harder (Gepstein and colleagues, per OR-07). Emma Hilton and Tommy Lundberg's review in *Sports Medicine* in 2021 draws the line across sport: a male advantage from puberty of 10 to 12 per cent in running, about 30 per cent in weightlifting at matched body mass, about 65 per cent in open powerlifting, and twelve months of testosterone suppression removing about 5 per cent of it (OR-07). Where the task is the body, two doors. Where the task is anything else, the evidence for two doors is thin: chapter 9 has the single-sex schooling meta-analysis and its rounding-error result, and Wong, Shi and Chen found in 2,059 school pupils and 456 students in 2018 that the one reliable product of a single-sex school was more anxiety around the other sex (OR-07).

But separate is not the same as subordinate, and the record's most durable arrangements kept the doors and shared the top: chapter 2's Haudenosaunee, whose 50 male chiefs were nominated and deposed by the Clan Mothers, and the Quakers' doubled plan of men's and women's business meetings, where Lucretia Mott learned to chair one (OR-06, OR-07). And there is a number for the share at which a minority starts to count. Bina Agarwal's 135 community forest institutions in Gujarat and Nepal, sampled in 2000 and 2001, show the all-women committees of Nepal handed the smaller forests, 20.9 hectares on average against 41.6, and regenerating them better, and show women's voice on a mixed committee beginning at roughly a quarter to a third of the seats (Agarwal 2009 and 2010, per OR-06). Below that, a woman on a committee is a woman in a room.

The kibbutz is the control. From 1910 the movement set out to abolish the sexual division of labour outright, with communal kitchens and children's houses, and by the third generation Lionel Tiger and Joseph Shepher's 1975 study found agriculture 88 per cent male, construction 99, management and politics 67, education 83 per cent female and childcare and services 85, and later work found women at 2 per cent of the senior central posts to men's 9 (OR-06). Tiger and Shepher called it biology. Their critics say the women were defined out of the fields by managers who called only heavy work productive, or that the return to the family was a revolt against an equality that meant living like men (OR-06). Either way the organisation that wrote no rule got the old one back.

Here is the tool, then. Every organisation answers two questions, and the record's failures answered them as one. Who does what is a question for the bodies and the bells: test the task and never the sex, and expect the sorting to be lopsided at the extremes and overlapping everywhere else. Who answers to whom is not a question for the bodies at all. It is a question for the rule, and if the rule does not answer it the room will, by whoever speaks most, and the room has a known accent. Robert of Arbrissel wrote the second answer down. We mostly let it emerge.

THE STRONGEST CASE AGAINST

The chapter has argued that an organisation should be written around the bodies it actually employs, and that the authority in it should be de-

cided and not left to emerge. Here is the best evidence that writing the body in is exactly the mistake, in its own numbers.

Every time a rule has named a female body, the owner of that body has paid for the naming. Jonathan Gruber's 1994 result is not a prediction, it is a measurement: mandated maternity benefits in American states came back to women of childbearing age as lower wages, nearly in full, because employers treated the mandate as a tax on hiring them (OR-04). Buriakova, Fitzpatrick and Machin found the same pattern across the OECD in 2022 for every sex-specific leave entitlement they could measure, with the hit landing hardest on women in small firms, precarious contracts and male trades (OR-04). Japan wrote menstrual leave into law in 1947 and 0.9 per cent of women touch it. Korea pays a day's extra wages to the women who refuse it, and about four in five refuse. The 955 young women who were asked preferred an explicit menstrual policy by 54.6 per cent (OR-04), and the economists' answer is that the cost of the explicit policy is in the offers they never see. On this reading the window and the fan in BS 30416 are harmless because nobody has to be named to use them, and everything in this chapter that names a body is a tax on it.

The second argument is that the body was never the reason. Galley's reading of the 1833 Factory Inquiry found women barred from the overseer's job, the best paid in the mill, long after steam had taken the strength out of it, which means strength was a word for a door, not a measurement of one (Galley 2013, *Continuity and Change*, per HC-04). Ogilvie's guilds re-passed their exclusions for eight centuries because the women kept qualifying. Hellum's Norwegian dormitories de-dramatised sex by removing the door, not by adding one. And the Marine Corps numbers, which are the chapter's strongest measurable case, are numbers about volunteers held to a standard built for men's bodies, in units that had been closed to women the year before, with a graduation rate that tells you about the standard as much as the women. A reader on this shelf would say the chapter has taken the ideal worker's rule, found that it hurts, and proposed a second rule written by the same hand.

The third argument comes from the other shelf and is shorter. Tiger and Shepher's 88 and 99 and 85 per cent say that when an organisation refuses to write the body in, the body writes itself in anyway, within three generations, in the most ideologically committed experiment ever run (OR-06). On that reading the abbess is not a design. She is a tail of the

distribution with a crown on, Hild a princess and the abbess of Fontevraud a figure for the Virgin, and a rule written for the tail is a rule nobody else can live under.

I think the first argument is right about labels and wrong about design, and the window is the proof. I think the second is right about the mill and the dormitory and does not reach the casualty carry. And I think the third describes exactly what happens when the two questions are answered as one. But the first is the one I cannot fully answer, because the firm that writes nothing down and the firm that names the body both have a bill, and the documents do not say which is larger.

IN PRACTICE

What follows is for a team, a department, a firm or a public body, whichever is yours, in conditionals, with the way each one fails.

If your organisation promotes the people who emerge, which means the people who spoke most in the meeting where no one had the chair, then the emergence data say you are selecting for the agentic accent and paying for it in the ratings afterwards. Decide the chair before the meeting and write the criteria for the job before the candidates are in the room. The failure mode is that the written criteria are themselves the leader stereotype in a numbered list, which Koenig's three paradigms say is still masculine, so read them aloud and ask which sex they describe.

If you are about to appoint a woman to a post that is already in trouble, then the FTSE data say you are about to do the usual thing, and the usual thing ends with the post and the woman both blamed. Give her the runway, the budget and the margin for error that the last man had. The failure mode is that nobody can tell, from inside the crisis, whether they are being brave or conventional, so look at who got the last three such posts in your own records.

If the work runs on long weeks and you are the senior man who sets the tone, then the cohorts say the week is the risk factor, not the stress you can feel, and the stigma data say the junior men are watching your hours and not your policy. Cut your own first, visibly. The failure mode is that a man given flexibility tends to work more to prove he did not want it, so the policy that lands is the one the most senior man is seen to use.

If you want to accommodate the cycle, pregnancy or the menopause, then the labelled leave data from Japan, Korea and the OECD say the label is a tax and the uptake is close to zero, while the window, the fan, the shift moved on request and the ten minutes out of the public eye are used, because nobody has to be named to use them. Build the second kind and make it available to everyone, which means the man with the migraine too. The failure mode is the register: the moment a firm keeps a list of its menopausal women it has a list, and Grandey's warning about who is experienced and who is symptomatic applies.

If you run a rota with fixed nights and a pregnant woman is on it, then the meta-analyses say a day role for the duration tends to lower a real risk, and the economics say that offering it to her alone tends to price her. Offer the move to any worker with a stated medical reason, at the same pay, and do not ask which reason. The failure mode is that a quiet offer is a quiet penalty, so the offer has to be the rule and not the exception.

If the task is the body, a casualty carry, a load over distance, a contact sport, then the Marine Corps and Caracal data say the sex gap at the extreme is large and does not wash out with training or with hormones, so test the task, at the standard the task needs, and let whoever passes pass. The failure mode cuts both ways: a standard set above the task's real need is Galley's overseer door with a stopwatch, and a standard quietly lowered for one sex is the Infantry Training Battalion's graduation rate waiting to happen in the field.

If you want women on a committee to change what the committee does, then Agarwal's forests say that below about a quarter of the seats they tend to be present and silent, and that the effect begins at a quarter to a third. One is a token. The failure mode is that the critical mass is a number from community forestry in Gujarat and Nepal, and your board is not a forest, so treat it as the lowest plausible figure and not as a target.

If your organisation has let its division of labour sort itself and finds, like the kibbutz, that it has sorted along the old lines, then the lesson is not that the lines were right. It is that the only question you forgot to answer in writing was who answers to whom. Answer that one on paper, as the abbey did, and let the other one sort. The failure mode is the oldest in the chapter: separate becomes subordinate the moment the authority is left to emerge, and the Second Council of Nicaea is still there to tell you which way it emerges.

And if none of this fits your organisation, then the overlap is doing its work. Every number above is an average, and the woman at the table on the fourth floor is one person, in one room, with one rule, which somebody, at some point, chose.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. The dual-hormone hypothesis as fallen, with Grebe 2019's power estimate and file drawer (OR-02). Eagly and Johnson 1990's near-zero style differences and Paustian-Underdahl 2014's self-versus-other split (OR-02). Wunder's working couple as the unwritten third design (OR-06). Badura 2018's two mechanisms, emergence by agency and rated performance by emotional reading (OR-02). Role congruity and the glass cliff as cultural layer (OR-02). Fontevraud's double rule under one abbess, 1101 to 1790 (OR-06, OR-07). Ogilvie's guilds as cartels re-passing their exclusions (OR-06, HC-04). Schoep 2019's presenteeism against absence (OR-04). Gruber 1994 on who pays for a labelled benefit (OR-04). Kivimäki 2015 and Pega 2021 on the 55-hour week (OR-05). The fatality ratio by sex (OR-05). MacCoun 1996's task-versus-social cohesion and the USMC 2015 task gaps (OR-07). Agarwal's critical mass (OR-06). Tiger and Shepher's snapback, with its critics (OR-06).

Decoration. The woman at the table and the man at hour fifty-six. Hild of Whitby. The Gilbertines' table. The Paris seamstresses' apprentice count. Lucretia Mott's chair.

Contested. Whether the dual-hormone hypothesis is dead (OR-02, OR-03) or alive (OR-05). Whether the Dekkers meta-analysis is Dekkers 2019 or Geniole 2016 (OR-02 against PS-05; chapter 10). Eagly and Johnson's study count, 162 against 161, and the 370 self-rating denominator (OR-02, PS-03). Whether the women-rated-better result is a filter or a bias (OR-02). Whether Huang 2026 is a paper or a press release (OR-05). How much authority the Clan Mothers held (Lafitau, Tooker, Viau, OR-06). Why the kibbutz re-sorted (Tiger and Shepher against their critics, OR-06). Whether the 33 per cent menstrual productivity loss exists on the employer's side (OR-04). Whether the preterm odds ratio of 1.21 is a 10 or a 21 per cent increase (OR-04, internal). Whether Hilton and Lundberg's baseline advantage is 10 to 50 or 30 to 80 per cent (OR-07, internal). Burnette's market against Galley's door (HC-04). Hartmann's pact against Humphries's room (HC-04).

NUMBERS TO CARRY

1. Leadership style: task-oriented $d = 0.00$, interpersonal 0.04, democratic against autocratic 0.22, 162 studies (Eagly and Johnson, *Psychological Bulletin*, 1990, per OR-02). Rated effectiveness: $d = -0.05$ overall, -0.12 favouring women when others rate, 99 samples, 101,676 people (Paustian-Underdahl, Walker and Woehr, *Journal of Applied Psychology*, 2014, per OR-02).

2. Dual-hormone interaction r = −0.061, 30 papers, 33 studies, 8,538 people (Dekkers et al., *Hormones and Behavior*, 2019, per OR-02); published literature at 16 per cent power, seven unpublished datasets null (Grebe et al. 2019, per OR-02).
3. Menstruation at work: 13.8 per cent absent, 1.3 days a year; 80.7 per cent present in pain, 33 per cent less output, 8.9 days a year lost; 32,748 Dutch women (Schoep et al., *BMJ Open*, 2019, per OR-04). Japan's menstrual leave used by 0.9 per cent (OR-04).
4. Fifty-five hours or more a week: stroke relative risk 1.33, CI 1.11 to 1.61, 528,908 people; heart disease 1.13, 603,838 people (Kivimäki et al., *The Lancet*, 2015). Attributable deaths in 2016: 745,194, 72 per cent male (Pega et al., *Environment International*, 2021). Both per OR-05.
5. Marine Corps task force 2015: 134 tasks, 93 significant differences, 88 favouring all-male teams; female Infantry Training Battalion graduation 36 to 46 per cent against about 98 (per OR-07). Fontevraud: 1101 to 1790, monks under an abbess (per OR-06).

THE LIBRARY

- OR-02 Leadership, hormones and sex: Mehta and Josephs, Dekkers, Grebe, Eagly and Johnson, Eagly 2003, Paustian-Underdahl, Badura, Eagly and Karau, Koenig, Eagly 1992, Ryan and Haslam, Croson and Gneezy, Garikipati, von Rueden.
- OR-04 Work design and female physiology: Schoep, the menstrual-leave table, Gruber, Buriakova, Dinour and Szaro, Wang, Bonde, Wallace, the Fawcett Society, BS 30416, Grandey.
- OR-05 Male work patterns and their costs: the dual-hormone hypothesis as live, Huang, Kivimäki, Pega, the Bureau of Labor Statistics, Milner, Pinquart and Schindler, Killewald, The New Dad, Rudman and Mescher, Farrell, Connell, Kimmel, Reeves.
- OR-06 Historical and traditional mixed organisations: Foot, McNamara, Nicaea II, Fontevraud, the Gilbertines, Vadstena, Ogilvie 2019, Bennett, Crowston, Wunder, De Moor and van Zanden, Agarwal, Rochdale, Raiffeisen, Tiger and Shepher, the Haudenosaunee.
- OR-07 Sex-segregated vs integrated institutions: Pahlke, Halpern, Wong, MacCoun, Hellum, the USMC task force, Gepstein, Hilton and Lundberg, Mahalik, Rotary v Duarte, the dual-governance table.
- HC-04 Industrialisation and the separation of home and work: Leyk, Burnette, Galley, Goldin, Horrell and Humphries, Davidoff and Hall, Hartmann, Humphries, Ogilvie 2003, Illich, Lewis, Correll.

CHAPTER 12

The Bus Route

Stand at a bus stop at ten past nine on a weekday morning, in any European city you like, and watch who gets on. The commuters have gone. The bus that comes now is half empty and the people who board it are carrying things: a pushchair with a toddler asleep in it and two bags of shopping hooked on the handles, a wheeled trolley, a man with a child's scooter under one arm and the child under the other. Nobody is going to the office. The woman with the pushchair rides four stops to the clinic, gets off, gets on again forty minutes later and rides three stops to the supermarket, then home on a transfer that has expired, so she pays a third time. The man with the scooter has dropped one child at nursery and is on his way to collect another from a grandmother. By lunchtime each of them has made four or five short trips that add up to a loop. The bus route under them is a line.

That line was drawn for someone. It was drawn for a passenger who leaves home at ten to eight, rides to the centre, works, and rides back at six, and it is a very good line for him. The timetable thickens at his hours and thins at everyone else's. The fare charges him once each way. The map is a star with the office district at its heart and the suburbs at the points, and the route from one suburb to the next, nursery to grandmother, goes through the centre or does not exist. He was, for most of the twentieth century, a man, and the people riding the loop at ten past nine were, for most of the twentieth century, women. I rode routes like his for years without once noticing who else was on the bus, which is the kind of thing a man born in 1975 can admit without much credit.

This chapter is about the largest thing human beings build and the fact that it was built, for the most part, with one passenger in mind. It is a short chapter in evidence terms and I want to say so at the start. One research document carries the city, and it leans, in places, on a magazine article and a blog. Where the chapter can stand on a number, it does. Where it stands on Vienna's own account of itself, I say so.

Ride the loop with them for an hour. Watch where the bus does not go.

Lines and loops

Transport surveys used to count two kinds of trip as real: going to work and going to school. Everything else went into a bin marked other, or escorting, or shopping, or leisure, and a bin is not a category anyone plans a bus route for. Inés Sánchez de Madariaga, an urban planner, gave the bin a name in 2009 and a framework in 2013, in a chapter of *Fair Shared Cities*, the Routledge volume on gender planning in Europe. She called it the mobility of care: all the travel done to keep a household running and its dependants alive, which is to say the nursery run, the clinic, the chemist, the shop, the grandmother, in whatever order the day demands.

Once the bin had a name, cities started counting what was in it. In Montréal, a regional travel survey reanalysed by Ravensbergen and colleagues in *Transport Policy* in 2023 found that caregiving accounted for 28 per cent of all trips, 32 per cent of women's daily trips and 25 per cent of men's. In Washington, Shuman and colleagues took the origin-and-destination records of the city's fare cards for the first quarter of 2019 and found that women were 58 per cent of all transit users, with a female share 10 to 15 percentage points higher on the off-peak, chained trips than elsewhere (Transportation Research Board, 2023). In Belo Horizonte a 2012 household survey, reanalysed in 2018, put care at about 15 per cent of daily commutes, 18 per cent of women's trips and 12 per cent of men's. In Bogotá the municipal mobility survey put it at 11.8 per cent of all trips, with the lowest-income women travelling most for care (Murillo-Munar and colleagues, 2023). The research document rounds all of this to somewhere between 15 and 30 per cent of urban travel.

Read those numbers for the overlap before you read them for the gap.

A quarter of men's trips in Montréal are care trips. The loop is not a female shape. It is the shape of whoever is doing the care, and the document says this plainly: a father who is the primary carer makes exactly the same chained, off-peak, short-hop pattern as a mother. The bus does not check your chromosomes. It checks whether you are pushing something.

So which layer is this? The loop is behaviour, and it is statistical: in every city measured, more of women's travel is care than of men's, by seven percentage points in Montréal and six in Belo Horizonte, and the margin is not a rule for the woman with the pushchair or the man with the scooter. What puts her on the loop more often than him is culture, the ledger of minutes that chapter 7 went through, and under that ledger sits a physiology, the gestation and lactation that chapter 8 priced, which starts the sorting before anyone has drawn a map. The route itself, though, is none of these. It is a decision.

The decision has a price, and the document names two mechanisms. First, fares. Most systems charge by the boarding, not by the hour, so a trip of four short hops costs four times a trip of one long one, and the person paying four times is the carer. Second, geometry. A hub-and-spoke network brings suburban workers to the centre and takes them home again. It does not run from one suburb to the next, which is where the nursery and the grandmother are. The carer who can afford a car buys one. The carer who cannot is, in the document's phrase, consigned to time poverty, and time poverty is the quiet tax on a loop that was never counted.

Who owns public space

In 1991 a planner in the Vienna city administration named Eva Kail put on a photography exhibition. Its title, in translation, was *Who Owns Public Space: Women's Everyday Life in the City*, and what it showed was the loop: women photographed through their days, in the stairwell with a pram, at the bus stop, in the park with the children, in the queue. It set off a public argument, which was the point. The year after, the city set up a Women's Office, and in 1998 it added a planning unit whose full name was the Co-ordination Office for Planning and Construction Geared to the Requirements of Daily Life, a title only a city administration could love and a function only a city administration could deliver.

In 1999 the city ran a survey of how its residents moved, and the survey drew the picture this chapter opened on. Men's trips were lines, home to work and back. Women's were loops, varied, local and tied to the small infrastructure of the neighbourhood: the nursery, the shop, the doctor, the stop. I have not seen the survey itself and the document describes it only through secondary accounts. But it is the origin of the governing image of this chapter, so I give it its place.

What Vienna built on that picture was not a theory of women. It was a procedure. The city called its method the four Rs: Representation, who sits at the planning table. Resources, where the money goes. Realia, what gets built. Rights, the zoning and the law. In 2013 Kail and Elisabeth Irschik wrote the city's *Manual for Gender Mainstreaming in Urban Planning and Urban Development*, which turned the four Rs into requirement profiles and quality indicators that every new development had to meet. The document counts more than 60 pilot projects over three decades. "Gender mainstreaming" is the term of art, and the word is the right one: the city sorted by gender, by who did what, and measured sex, the body, only where the body makes its own demands, which we come to at the toilet door.

Two of the projects carry the chapter. The first is housing. In the late 1990s the architect Franziska Ullmann and others designed Frauen-Werk-Stadt, Women-Work-City, in two phases: 359 subsidised flats with a kindergarten, a medical centre and shops built into the site, so that the loop could be walked rather than ridden. The flats were arranged around grassed courtyards so that a child playing could be seen from the kitchen window. Stairs and entrances were widened and lit so that there were no dark corners and a pram could get up them. If you have read chapter 2 you recognise the shape. A foraging camp is a clearing where the children are visible from where the adults work, and where any adult who looks up is, for that moment, on duty. The courtyard is the Campfire Contract poured in concrete, and the kitchen window is the fire.

The scheme was scaled up at Aspern Seestadt, a brownfield expansion on the city's edge planned for 20,000 residents and 20,000 jobs by 2028. There the city wrote the requirements into the developers' competitions. A firm that did not provide pram storage, communal rooms and a safe lighting plan was denied the public subsidy and the contract. In 2024 an evaluation report for Aspern found that 93 per cent of residents rated their

quality of life highly, which I pass on with the warning that the document's source for it is the developer's own press release.

The second project is a park. Kail and the landscape architect Claudia Prinz-Brandenburg had watched Vienna's parks and found that girls disappear from them after about the age of nine, while boys of the same age become the dominant users. The reason, in the workshops, was the cage: the fenced football or basketball pitch with one entrance, which a group of older boys could hold as territory and which a girl could enter only by walking into a space she could not easily leave. The girls said so. They did not want to be shut in there with boys they did not know, and when they tried to share the pitch they were shouted at or worse. At Einsiedlerpark in the fifth district the city opened the fences on several sides, widened the paths, improved the lighting and spread the things girls had asked for along the edges of the park where you could reach them without crossing the pitch: hammocks, seating platforms, a volleyball net, badminton courts. Girls came back, and stayed longer, and from 2005 the lesson became a binding guideline for every park the city designs. I should say that the "after nine" figure has no primary source in the document, which cites a magazine article and a blog case study, and that the outcome is reported by the city about itself.

Notice what the park did not do. It did not decide that girls are timid or that boys are territorial. It observed who held the space and who left it, and it changed the fence. That is the whole of the Vienna method and it is why a chapter on cities can live, a little uneasily, in a book about hormones. The accent, whatever it is, is heard in the park. The fence is ours.

The queue

Now the body, because there is one place where the city measures sex and not gender, and it is the lavatory.

Clara Greed, writing in the *Town Planning Review* in 2016 and again in 2021, has spent a career insisting that public toilets are a planning question, not a plumbing one, because they decide who can be out of the house for how long. Building codes have historically answered the question with spatial equality: the same floor area for men and for women. The result is a queue outside one door and none outside the other, and in 2020 Bovens

and Marcoci put the queue into a model, in *Behavioural Public Policy*, to show why. The average service time for a woman is 178.9 seconds and for a man 118.4. That is a measurable fact about bodies and their circumstances, and the circumstances come from all three layers: menstruation and pregnancy are physiology, the clothing is culture, and the child or elderly relative who has to be helped is the loop again, following the carer into the cubicle. On top of the time there is the hardware. A men's room of the same floor area holds more fixtures, because urinals pack tighter than cubicles. And queues do not grow in proportion to the difference that feeds them. Near capacity they grow much faster than that, so a service time half as long again, plus fewer fixtures, produces a queue out of the door on one side and an empty room on the other.

Every woman who has waited through the interval while the men's door swung freely already knew the finding. It took a queuing model to make it a planning issue.

The model offers two remedies. One is to abandon equal floor space and mandate a fixture ratio in women's favour, which the literature puts between 2:1 and 3.25:1. The other is to take the sign off the door. Bovens and Marcoci's simulations, run at different intensities of use, found that a desegregated room of enclosed cubicles always shortened women's waits, and shortened or only slightly lengthened men's, depending on whether urinals survived the redesign. The document notes that the public argument about shared toilets has been about trans people, and that the queuing result is simply about throughput for everyone. It also notes, under its open questions, that whether people accept the shared room varies a great deal from one culture to another, so the model is not the building.

The rest of the body's claims on the city are smaller, and the document gives no primary source for any of them, so I leave them out. What it does give is a finding about what a city's shape does to activity. Density moves everyone: in the IPEN Adult study, 11,615 people in 14 cities (Giles-Corti and colleagues, *The Lancet Global Health*, 2022), the dense, well-connected, well-served neighbourhoods were the ones where people reached the 150 minutes a week that health guidelines ask for. The sex difference sits one level down, in a 2020 systematic review by the same group in the *Journal of Epidemiology and Community Health*: women's activity responded more to good public transport, physically separated cycle lanes and having daily destinations close by, men's more to raw street con-

nectivity and the absence of dead ends. That is a statistical finding about overlapping populations, not a different species of pedestrian. But it is the difference between a plan that draws a dense grid and calls the job done, and one that also builds the kerb cut and the lane with a barrier.

After dark

The third layer is the night, and here the evidence turns against the chapter, so I want to give it room.

Women report more fear in cities after dark than men do, consistently, and the document is careful to say that this fear is not an opinion. It is a restriction: it alters which routes are taken, which transport is used and whether a person goes out at all, and it does this whether or not the objective risk matches it. Every city's answer to the fear is the same answer, which is more light, and the question of whether light reduces crime is one of the longest-running arguments in criminology.

The case for the lamp is a Campbell systematic review by Brandon Welsh and David Farrington, criminologists, published in 2008 as an update of their 2002 analysis. Thirteen evaluation studies from Britain and the United States, and improved street lighting reduced crime by 20 per cent in the lit areas against controls, an odds ratio of 1.27 with a confidence interval of 1.09 to 1.47. Then the odd detail. Night-time crime did not fall more than daytime crime. Welsh and Farrington took this to mean the mechanism was not surveillance, not that the offender could be seen, but something they called community pride: the city visibly spending money on a street, and the street's residents answering with more order of their own. The lamp as a signal, not a floodlight.

The case against has two parts. Paul Marchant, a statistician, argued in a series of critiques from 2004 to 2013 that the review's arithmetic was wrong at the root. Crime is clumpy, not evenly scattered, so the usual error bars are too narrow. And lights get installed where crime has just spiked, so a fall afterwards is what you would see anyway as the spike subsides. On that reading the 20 per cent is an artefact of where councils put lamps. Then, in 2015, Steinbach and colleagues published in the *Journal of Epidemiology and Community Health* an interrupted time-series analysis across 62 local authorities in England and Wales over 14 years, from 2000

to 2013, a period in which councils had been switching lights off, dimming them, running them for part of the night or replacing them with white LEDs to save money and carbon. For the offences most likely to happen at night, burglary, vehicle theft, robbery, violence and sexual assault, they found no association between the lighting changes and crime, and none with night-time traffic collisions either.

The document's reconciliation is that the two findings are about different scales: a lamp in one bad alley may shift what happens in that alley, and may shift it to the next one, while the lighting level of a whole county does not move its crime rate. I find that plausible and I notice that it is also convenient. What survives either way is the fear, and the planner's move, which I think is the right one, is to separate the two questions. Light is a contested tool against crime. It is an uncontested tool against fear, and fear is what keeps a woman off the last bus. The three-layer sentence here is awkward and I will say it anyway: the fear is cultural in origin, statistical in its distribution between the sexes, and its effect on who is at the stop at eleven at night is as real as the timetable.

The lonely city

A city does one more thing besides moving people and lighting them. It introduces them. For most of human history, as the second research document puts it, mates were brokered by the people around you: family, friends, the congregation, the street, who did the vetting by knowing everyone. The camp that chapter 2 described worked the same way. The city replaced the camp with the neighbourhood, and the neighbourhood is now being replaced by something else.

Start with who is alone. In 2024 about 17 million people in Germany lived in single-person households, 42.1 per cent of all households, well above the European average (Destatis, 2024). Being alone is not the same as being lonely, and the German loneliness barometer of 2024, run on the Socio-Economic Panel's 2021 and 2022 waves, found that 19 per cent of the population feels lonely at least sometimes. Among them, 20 per cent describe isolation, 28 per cent exclusion and 56 per cent simply missing company. The groups most at risk are people below the median income, men, and people with a migration background, and the predictive weight

of loneliness falls, against the folklore, on the young and the middle-aged rather than the old (Luhmann and colleagues, on the same panel and the Berlin Aging Study II). In the German family panel, 7,165 non-cohabiting relationships over 13 waves, young adults who are content being single are fine, and those who want a partner and have not found one decline in life satisfaction, and keep declining.

Now how people meet. Rosenfeld, Thomas and Hausen, using a representative American survey of how couples met, showed in *PNAS* in 2019 that meeting online overtook meeting through friends around 2013 and that by 2017 about 39 per cent of heterosexual couples had met that way. Finkel and colleagues, in a 2012 review in *Psychological Science in the Public Interest*, described what the screen does to the person looking at it: it puts them in an assessing frame of mind, comparing profiles the way one compares flats, and people who compare commit less.

Then the result that I think matters most for a city. In 2017 Joel, Eastwick and Finkel, in *Psychological Science*, took two speed-dating samples, 163 and 187 people, gave each of them more than a hundred questionnaire measures of the traits and preferences the literature says matter for mate choice, and set a machine-learning model to predict who would want whom, tested on people the model had not seen. It could predict 4 to 18 per cent of how much a person liked people in general, and 7 to 27 per cent of how much people in general liked them. Of the unique desire between two particular people it predicted 0 per cent. Nothing. Which means that the thing the apps filter on, stated traits, is the thing that does not predict the only outcome anyone wants, and the thing that does produce it, two people in the same place by accident, is what a bus stop, a park and a courtyard are for.

Before the chapter sounds like it is about women, the other column. In the American General Social Survey from 2000 to 2018, 4,291 men and 5,213 women, Ueda and colleagues found in *JAMA Network Open* in 2020 that the share of men aged 18 to 24 reporting no sex at all in the past year rose from 18.9 per cent to 30.9, an age-adjusted odds ratio for the trend of 1.20 with an interval of 1.04 to 1.39. The rise was not evenly spread. A man earning 50,000 dollars or more had an odds ratio of 0.37 for inactivity against a man earning under 10,000. Part-time work carried 2.08 and no work 2.94. Money, though, does not buy the old arrangement back. When fracking raised the wages of men without a degree in particular

American districts, Kearney and Wilson showed in the *Review of Economics and Statistics* in 2018 that births rose, inside and outside marriage, and marriages did not. The Appalachian coal boom of the 1970s and 1980s had raised both. At the far margin, Costello and colleagues profiled 151 self-described involuntary celibates against 149 other single men in 2022: 74.1 per cent of the first group reported severe anxiety and 67.5 per cent severe depression, and, against the caricature, they reported lower minimum standards for a partner than the comparison men, not higher. The loneliest people in the city are men on low incomes, which is not what a chapter about building cities for women set out to find, and I would rather say it than tidy it away.

Two more findings pull in the same direction. Couples in which the wife has more education than the husband now outnumber the reverse across Western countries, and they are no longer less stable (Van Bavel, Schwartz and Esteve, *Annual Review of Sociology*, 2018). And when Eastwick and Finkel measured in 2008 what people actually pursued at speed-dating events against what they said they wanted, the sex difference in stated preferences, looks for him, earnings for her, disappeared in the room, a result a worldwide replication confirmed in 2024. The mating market's trouble, on this evidence, is not that men and women want different things from each other. It is that the places where they used to find out have been replaced by a screen that cannot do the job.

A city for the loop

So what would a city built for the Campfire Contract look like? Not, I think, like a city built for women, and the first document agrees: to design a city that fixes women as the carers because they are, statistically, the carers, is the naturalistic fallacy with a planning permit. The aim is to take the friction out of the loop for whoever is walking it, so that the question of who walks it can be decided at the kitchen table and not by the timetable.

The nearest existing proposal is the fifteen-minute city, the neighbourhood in which work, shops, the doctor and the school are a quarter of an hour's walk or ride away, and it fits the loop well. Leslie Kern, a feminist urban geographer, warned in *Feminist City* in 2020 what happens if

the market builds it: the grocer, the nursery and the café go upmarket, the low-income women who do most of the city's paid care are priced out to the edge, and then commute back in to do it. Proximity as a luxury good undoes itself. Her conditions are social housing, rules against displacement and care provided collectively, which is to say Vienna's conditions, and Vienna has them because it funds housing. Berlin's building groups, the Baugruppen, are the small-scale version: residents form a legal entity, commission their own block, and come in 25 to 35 per cent under the market, and almost always build in a shared kitchen, a garden, a childcare room and a guest room, so that care is done among neighbours rather than alone. The failure mode is in the first clause. People who can form a legal entity and finance a building are not, on the whole, the people Kern is worried about.

One last thing, borrowed from the third document, whose real home is chapter 15. Infrastructure has sorted the sexes before. The plough needed upper-body strength and could not be worked with an infant on the hip, and Alesina, Giuliano and Nunn showed in the *Quarterly Journal of Economics* in 2013, with the Ethnographic Atlas and an instrument built from which crops a society's land favoured, that the descendants of plough cultures still have fewer women in work, in business and in politics today, and that the pattern follows the children of migrants into countries that never used one. The tool is gone. The sorting outlived it by centuries. A city is a tool too, and at Aspern 20,000 people are about to move into one. Whichever passenger the route is drawn for is the passenger the route keeps producing.

So the three questions, at this scale. What is measurable: the queue, the pram on the stair, the 178.9 seconds. What is statistical: the loop, its 28 per cent, its 32 against 25, and the lonely man on a low income whom nobody's plan was drawn for. What is ours: whom the line is drawn for. The route is a line somebody drew. Here is the pencil.

THE STRONGEST CASE AGAINST

The chapter argued that cities were drawn for one passenger, that the frictions fall on the carer and that changing the fence, the fare and the lamp lowers them. Here is the best case that it has argued from one city's self-portrait and a handful of surveys.

Start with the lamp, because it is the intervention every city reaches for first and it is the one with a failed replication. Steinbach and colleagues, 62 local authorities, 14 years, every kind of lighting cut, and no association with burglary, robbery, violence or sexual assault (*Journal of Epidemiology and Community Health*, 2015). Marchant's critiques say the 20 per cent in Welsh and Farrington's review is what regression to the mean looks like when councils light the streets that just had a bad year. The chapter's reply was to retreat from crime to fear. That is honest, and it is also a retreat: fear is measured by asking, and a city that lights a street and then asks its women whether they feel safer is marking its own homework.

Then Vienna. The entire city-scale case rests on one document, and that document's sources for its best scenes are a magazine feature, a blog case study and a developer's press release. The 1999 survey is described second-hand. The park outcome is the city's report on itself. The 93 per cent is a satisfaction poll of people who chose to live there. No one has compared the women of Vienna with the women of a city that did nothing and found that they travel faster, work more or fear less. Thirty years of gender mainstreaming may have done a great deal. On the evidence in front of me, it has also not been tested.

Third, the loop is not a sex difference at all, and the chapter half admitted it. Hyde's 2005 synthesis in *American Psychologist* found 78 per cent of psychological sex differences near zero or small and a typical gap of 0.21 of a standard deviation, an overlap above 85 per cent, and nothing in the travel surveys suggests the city is an exception: a quarter of men's trips in Montréal are care trips, and the father with the scooter makes the mother's loop. If the loop is a carer's loop, then the chapter is about care, not about sex, and a book about the Endocrine Accent has wandered into someone else's subject. And the first document's own warning cuts against its own flagship. A city that builds the loop into its fares and its courtyards for the people who currently walk it is a city that makes the loop comfortable, and comfortable arrangements persist. The plough is the proof.

Fourth, the lonely city. The people the chapter found most alone are men on low incomes, and nothing in the planning literature touches them. Ueda's odds ratios, Kearney and Wilson's fracking districts and Costello's 151 are a story about wages, work and a mating market that money no longer converts into marriage, and no fence, fare or lamp is ad-

dressed to it. A chapter that opens with a woman on a bus and ends with a man on a low income has changed the subject, and should say which one it is about.

Last, the sexes may want the same city. Eastwick and Finkel's revealed preferences were the same in the room, in 2008 and in the 2024 replication, and Van Bavel's wives who out-educate their husbands are not divorcing more. If the role the chapter planned for is dissolving under its own evidence, then the right city is the one that is good for a person with a pushchair, whoever they are, and the word "women" in Vienna's manual is already out of date.

IN PRACTICE

If you sit on a council, draw a timetable, run a housing office or only ride the bus, this is what the evidence supports at the city's scale, in conditionals, with the places it fails.

If your travel survey files the nursery run under "other", then you do not know who rides your network and you cannot plan for them, so count the loop first, as Montréal and Washington did, before you change a single route. The failure mode is that counting is not building: Montréal's 28 per cent is one survey in one city, and a figure in a report tends to stay in the report.

If your fares charge by the boarding, then a four-hop morning costs four times a one-hop commute and the person paying four times is the carer, so charge by time and let the transfer cover the loop. The failure mode is revenue and fraud control, which per-boarding fares also do, and the evidence is an inference from fare-card data (Shuman and colleagues, 2023), not a trial of the alternative.

If your network is a star, then the trip from one suburb to the next goes through the centre or not at all, so run the orbital route the survey shows and accept that it loads lightly. The failure mode is the empty bus, which a transport authority tends to read as a failed route rather than a served one.

If your building code gives the sexes equal floor space for toilets, then the women's queue is a mathematical certainty at 178.9 seconds against 118.4, so either set the fixture ratio at 2:1 or above or take the sign off a room of enclosed cubicles. The failure mode is acceptance: the simulation

is a model, the document's own open question is whether people in your city use a shared room at all, and men's waits can lengthen slightly if the urinals go.

If your parks have a fenced pitch with one gate, then the pitch tends to belong to whoever holds it, and the evidence from Vienna is that the ones who leave are girls from about nine. So open the fence on several sides and put the things they asked for along the edges, as Einsiedlerpark did. The failure mode is that the boys lose a pitch and say so, and that the evidence is a case study on a blog with no primary evaluation behind it.

If you light a street, then say which problem you are lighting it for. For fear, the evidence is consistent and the fear is a real restriction on who goes out. For crime, the evidence is Welsh and Farrington's 20 per cent against Steinbach's nothing across 62 authorities, and a council that promises a crime reduction tends to be caught by its own statistics the following year.

If you subsidise housing, then make the subsidy conditional, as Vienna did at Aspern: pram storage, a lit stair without corners, a courtyard the kitchen looks onto, childcare on the ground floor. The failure mode is Kern's: in a market without social housing the same features price out the carers they were built for, and the 93 per cent satisfaction is a press release, not a comparison.

If you want a fifteen-minute city, then write the rules against displacement before the café opens, not after. The failure mode is that proximity becomes a luxury good and the care workers commute in from the edge to serve it, which is the old line with a new name.

If people in your city are forming building groups, then make the legal and financial path short, because the shared kitchen, garden and childcare room appear almost every time a group designs its own block. The failure mode is in who can form a group: 25 to 35 per cent under the market is a saving for people who could have bought at the market.

If you worry about loneliness, then notice who is lonely in your data before you plan for them. In Germany it is men on low incomes and people with a migration background, young and middle-aged more than old, and the planning literature on cities for women has nothing addressed to them. The thing a city can do that an app cannot is put two people in the same place by accident, which is the one thing the machine-learning models could not predict at all. So build the places where accidents happen, the bench, the park edge, the courtyard, and expect noth-

ing in particular from any one of them. The failure mode is that nobody has run that trial, and that the lonely may not come.

And if the loop in your city is walked mostly by women, then build for the loop and not for the women, so that whoever walks it tomorrow finds the route already there. The failure mode is the plough: a city that makes one arrangement comfortable tends to keep it, and the person across the kitchen table may have wanted a different one.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. The mobility-of-care surveys: Montréal 28 per cent, 32 against 25 (Ravensbergen 2023), Washington's 58 per cent and the off-peak female share (Shuman 2023), Bogotá's 11.8 per cent (Murillo-Munar 2023) (SO-03). Bovens and Marcoci 2020 on service times and queue growth (SO-03). Vienna's institutional record: 1991 exhibition, 1992 office, 1998 unit, 1999 survey, 2005 park guidelines, 2013 manual (SO-03, from municipal and secondary sources). Welsh and Farrington 2008 and Steinbach 2015 carried as a pair (SO-03). Destatis 2024 and the loneliness barometer 2024 (SO-06). Rosenfeld 2019 and Joel 2017 (SO-06). Ueda 2020 and Kearney and Wilson 2018 (SO-06). Alesina, Giuliano and Nunn 2013, borrowed from chapter 15's material (SO-05). Said plainly: one document, SO-03, carries the whole city-scale argument, and it leans on a magazine, a blog and a press release for its best scenes. The chapter is modest because the library is.

Decoration. The bus stop at ten past nine. The kitchen window as the fire. The hammocks. The interval queue. The pencil.

Contested. Whether lighting reduces crime (Welsh and Farrington against Marchant and Steinbach). Whether the Einsiedlerpark result and the "after nine" figure have a primary source. The 93 per cent at Aspern (press release). Acceptance of desegregated toilets (SO-03's own open question). Whether the fifteen-minute city can be built by the market (Kern). Whether the chapter is about sex at all, or about care.

NUMBERS TO CARRY

1. Service time at the toilet: 178.9 seconds for women against 118.4 for men, with a fixture ratio of 2:1 to 3.25:1 needed to equalise the queue (Bovens and Marcoci 2020, *Behavioural Public Policy*, per SO-03).
2. Care trips in Montréal: 28 per cent of all trips, 32 per cent of women's and 25 per cent of men's (Ravensbergen and colleagues 2023, *Transport Policy*, per SO-03).
3. Street lighting and crime: 20 per cent reduction, odds ratio 1.27 (1.09 to 1.47), 13 studies (Welsh and Farrington 2008, Campbell Systematic Reviews), against no association across 62 English and Welsh authorities over 2000 to 2013 (Steinbach

and colleagues 2015, *Journal of Epidemiology and Community Health*), both per SO-03.

4. Germany 2024: 42.1 per cent of households are one person, about 17 million people (Destatis), and 19 per cent of the population lonely at least sometimes, highest among low-income men and people with a migration background (Einsamkeitsbarometer 2024, SOEP 2021/22), per SO-06.
5. Predicting desire: 4 to 18 per cent of a person's general liking, 7 to 27 per cent of their general desirability, 0 per cent of the unique pair, 163 and 187 people (Joel, Eastwick and Finkel 2017, *Psychological Science*, per the SO-06 PDF). Men aged 18 to 24 with no sex in the past year: 18.9 to 30.9 per cent, 2000 to 2002 against 2016 to 2018 (Ueda and colleagues 2020, *JAMA Network Open*, per SO-06).

THE LIBRARY

- SO-03, Cities designed for both sexes: the whole city-scale argument, Vienna, the mobility of care, the toilet queue, the lighting controversy, Kern and the Baugruppen.
- SO-06, The contemporary mating market and loneliness: the German loneliness figures, how couples meet, the predictability of desire, the retreat of young men, taken only where it touches the city.
- SO-05, Cultural evolution: why patriarchy arose, persists and weakens: the plough as the precedent for infrastructure that sorts the sexes, one paragraph only. The rest is chapter 15's.

CHAPTER 13

The Wall

It is a quarter to one on a weekday in a town in the west of Germany, and a woman is standing at a school gate with her car keys in one hand and nothing in the other. In fifteen minutes the doors open and the day's schooling is over. Not the lesson. The day. She has arranged her life around it the way you arrange a kitchen around the position of the sink, which is to say without ever quite deciding to. Behind her, in the car park, are other women with other keys. There is one man, and he is, in the way of these things, a little conspicuous.

That evening she sits at the kitchen table with a tax form. The form asks for her husband's income and for hers, adds them together, divides the sum by two, looks up the tax on one half, and doubles it. Four operations. They were written into West German law in 1958, the year my mother turned twelve, and they are still there, which means they outlived her. What they do to her first hour of paid work comes below. Her hours, in any case, have already been set by the school.

Now come with me, because this is the chapter where the book goes to the one place on earth that ran the experiment properly.

On the other side of a border that no longer exists, inside the same language and the same Sunday cake, another woman of the same age grew up in a country that took the opposite bet on every one of these questions, and took it for forty-one years. The crèche was down the road and it was open all day. The tax office did not care who in the household earned the money. Nobody at the gate at a quarter to one, because there was no gate at a quarter to one. And the word in the air for a woman who stayed at

home was not raven mother. It was parasite.

Then in 1990 the border came down and the western rules were laid over the whole country at once: the same tax form, the same timetable, the same word for the woman at the gate. If behaviour followed the rules, the two women should have become one woman within a few years. The social-security records say they did not. They say the eastern mother is still at work, a generation on, for about a month and a half more a year than the western one.

There was a wall. It was made of concrete, and it was also made of a timetable and a tax table, and only one of those three came down.

The paragraph and the cover

The Preface told you about the cover. Let me tell you what was on the other side of it.

Until the late 1970s the legal name for a West German marriage was the *Hausfrauenehe*, the housewife marriage. A wife could take a job only with her husband's consent, and the consent could be withdrawn if the job got in the way of what the law called her duties in the home (HC-05, which names no statute, as the Preface already notes). Abortion was a crime under Paragraph 218 of the criminal code, with prison as the sentence. These were the two walls the movement my mother's generation joined could see, and it went at the second one first, because the second one was about the body and the body was where the argument could not be lost.

Alice Schwarzer, a journalist who had seen the French Manifeste des 343 that spring, gathered 374 women, Romy Schneider among them, and put their faces and their confession on the cover of *Stern* on 6 June 1971. It made her the face of West German feminism and led, a few years later, to the magazine Emma. In 1974 the social-liberal government passed the Fristenlösung, a law making abortion legal in the first three months. In 1975 the Federal Constitutional Court struck it down (BVerfGE 39, 1), reading into Article 2 of the Basic Law a right to life for the unborn that the state was bound to protect, against the mother if need be, and the country fell back on a system of indications: abortion allowed for medical or criminal reasons and not otherwise. The German Democratic Re-

public, three years earlier, had legalised abortion on demand in the first twelve weeks. The compromise of 1993, after reunification, is the one still in force: abortion remains an offence under Paragraph 218 and goes unpunished in the first twelve weeks if the woman has been counselled and has waited (HC-05).

Count what was won. The right to the body, in the compromised form a court allowed. The partnership marriage, in 1977, when the housewife marriage was struck from the law and a wife no longer needed anyone's permission to work. They were fought for against the memory of a state that had treated women as a means of producing children, and it is that memory which explains the movement's one great blind spot.

West German feminists did not, on the whole, demand public childcare. Many of them distrusted it. A crèche run by the state looked, from 1971, like something the previous regime would have built, or the one across the border was building, a place where children were collectivised and mothers made interchangeable (SY-06, which sources the claim to a literary study and a paper on motherhood, not to the movement's own documents). So the movement fought for the right to work and not for the means to. It won the right. And a woman who used it found herself, "entirely on an individual basis" against a tax code, a timetable and a word, with no feminist vanguard behind her on the one question that decided her hours (SY-06). Emancipation, in the West, became the double burden: a part-time job in the morning and the whole of the house for the rest of the day, because the men of the house had not been part of the argument and did not change.

That is the culture layer, and none of it was German in origin. The idea that a home is a sphere apart from work, run by a woman whose idleness proves her husband's standing, was written in English in Birmingham and Essex between 1780 and 1850, in sermons, wills and house plans (Leonore Davidoff and Catherine Hall, *Family Fortunes*, 1987, per HC-04). The model travelled. When Jane Lewis sorted the welfare states of Europe in 1992 by how far their tax, pension and childcare rules assumed a wife at home, she put Britain and Germany in the strong male-breadwinner column and Sweden and Denmark in the weak one (Lewis, Journal of European Social Policy, 1992, per HC-04). Germany had not invented the breadwinner. It had written him into the tax code more carefully than anyone else.

Pool, halve, tax, double

Let me pay for the four operations, because the arithmetic is the mechanism and the mechanism is the chapter.

Take a couple in which he earns a great deal and she earns nothing. Pool the incomes, halve the pool, and the tax schedule now sees two moderate earners instead of one large one. Because the schedule is progressive, two moderate tax bills come to less than one large one, and the saving, the splitting advantage, is largest exactly when the gap between the two incomes is largest, which is to say when one of them is zero (SY-06). Now let her take a job. Every euro she earns is added to a pool that is already sitting at his marginal rate, so her first euro is taxed as if it were his last. The couple's advantage shrinks with every hour she works. A microsimulation by Viktor Steiner and Katharina Wrohlich in 2004 found this one rule to be a primary driver of the exceptionally low full-time employment of married women in Germany (SY-06, which places the study at CESifo while its own works-cited list gives a DIW discussion paper).

Nobody is forced to do anything by this. That is the point. A tax code of this kind does not forbid. It prices. It prices her hour of work below his, in a household that has every reason to count hours in money, and it does so in a schedule that neither of them wrote and few of them can explain. Add the timetable: schools that ended at lunch, and, for the under-threes, public care that was for practical purposes a rumour. The state's own infrastructure assumed somebody would be at home in the afternoon, and so it required it (SY-06). The two rules did not need each other. They fitted, the way a sink fits a kitchen.

Then the word. A *Rabenmutter* is a raven mother, cold, a bird that leaves its young in the nest, and the word did what the tax code and the timetable could not: it turned a logistical problem into a defect of character. A woman who put a two-year-old in care was not short of options. She was short of love. My document gives no history for the term and takes its definition from a forum thread. But the economists who later measured the stigma did not need the etymology. They measured the hours.

A word about layers, because this is where the book's rule earns its keep. The tax code is culture: it is variable, it was voted on, it could be voted away. Its effect is statistical: not every married woman in the West stayed home, and the ones who did not were paying a price, not breaking a

law. And the thing being taxed was a body: the one body in the household that could gestate and lactate, and which the state, with the best Catholic intentions and the worst arithmetic, decided to shield from the market by making the market expensive for it (SY-06).

The compounding is the part my parents' generation did not see. Chapter 7 carries the German pension figures: older women drawing about half the pension income of older men, because the pension follows lifetime earnings and lifetime earnings follow the hours dropped at the first birth (CO-02, via chapter 7). The welfare state that was built to protect the family from the market protects it, in the end, by making the mother dependent on it at seventy. Nobody designed that either.

Mummy politics

The East did not start from a theory of the family. It started from a shortage of hands.

The German Democratic Republic lost workers to the war and then to the border, which was open until 1961, and a planned economy that cannot find labour has an interest in women that a market economy can afford to lose. The constitution of 1949 wrote in the right to work and to equal pay. Public childcare was built, and by the late 1980s it covered, in SY-06's phrase, nearly every child in the country, so that the logistical barrier to a mother's employment did not so much fall as fail to exist (SY-06, HC-06). The result was the figure the Preface gave you: by 1989 nine women of working age in ten were in the labour force in the East, against a little over half in the West (Rachel Rosenfeld, Heike Trappe and Janet Gornick, *Annual Review of Sociology*, 2004, per HC-06).

Then the planners noticed something a planner would notice. The women were working and they were not having children. Fertility fell, and from the 1970s the state turned to what Germans on both sides later called Muttipolitik, mummy politics: policy designed not to free women from motherhood but to make it compatible with the plan (SY-06). The flagship was the *Babyjahr* of 1976, a paid year off after a birth, the job held open, with the option of reduced hours afterwards and no penalty on the way back (SY-06, which names no primary source for the 1976 provisions). Note the name. It was mother politics. The year went to her. The reduced

hours went to her. The state that had abolished the breadwinner's tax advantage had not abolished the idea that the baby was hers, and that is going to matter in a few pages.

And the word ran the other way. A woman who stayed home with her children was a *Schmarotzer*, a parasite on the collective, somebody avoiding her civic duty (SY-06, same forum source as above). Two states, one language, and each had found a single noun to put on the woman who did what the other state required.

Inside the flat, nothing like this symmetry held. The East put women into the shift and the office and left the kitchen to them as well. Trappe and Rosenfeld documented the result across the GDR's own surveys: full-time paid work plus near-total responsibility for the home, the double burden, and men whose domestic hours did not move to meet their wives' paid ones (HC-06). HC-05 puts it more sharply: the socialist state did not challenge male privilege at home. The emancipation of East German women was a labour policy, delivered from above, and a labour policy does not need to redecorate the kitchen.

So when the two movements met, in the early 1990s, each saw a stranger. Western feminists looked at the East's Independent Women's Association and saw obedient subjects of an authoritarian state. Eastern women looked back and saw people who had argued about the body for twenty years while they ran a household on a full-time wage, and who now stood by while the crèches and the *Babyjahr* were dismantled (SY-06, sourced to a book preview). Both were right about the other. Neither had the other half.

What the records say

It is now thirty-five years since the rules were made the same, and the best part of the experiment is that German bureaucracy kept the receipts.

Barbara Boelmann, Anna Raute and Uta Schönberg took a 50 per cent random sample of the country's social-security records and built from them what economists call a collective model of family labour supply: in which the couple, not the individual, decides who works how much, and the price of a mother's hour includes the social cost of being seen to work (SY-06). The Preface gave you the headline. The mechanism is the

part that belongs here. An East German mother works about 1,800 hours a year. A West German mother about 350 fewer. Fathers, East and West, the same (Boelmann, Raute and Schönberg, *American Economic Journal: Applied Economics*, 2025, per SY-06).

Now the decomposition. Wages are lower in the East and childcare is dearer in the West, so a naive reading says the East works more because it is paid less and charged less. The model says the opposite. Equalise the wages across the country and the gap does not close. It nearly doubles. The higher West German wage is, in effect, what the market pays a western mother for working against the word, and take it away and more of her stays home (SY-06). The wage was holding the western figure *up*. Their estimate, repeated here because it is the hinge of the chapter, is that the working-mother stigma alone accounts for 73 per cent of the gap.

Then the movers, and what the Preface left out. Not only did West German mothers who moved east come to work like East German mothers, but the ones who later moved back west kept doing it (SY-06). The behaviour did not depend on the crèche being outside the door. It had been learned and it travelled. And in West German firms where the share of East German colleagues rose, western mothers went back to work after a birth sooner (SY-06). You can catch the East at the office.

Henrik Kleven and his colleagues had already shown in 2019 where Germany sits in the world. They follow a woman's earnings through her first birth against a man's, and call the long-run gap the child penalty. Ten years on, a West German mother is earning about 61 per cent less than she would have been, or 60 by another of my documents, against about 20 per cent in the United States and 21 to 27 in Scandinavia (Kleven et al., *AEA Papers and Proceedings*, 2019, per HC-05, with chapter 8 carrying the 60-to-61 range). In a 2024 working paper Kleven turned the same instrument on the inner-German border and found that mothers raised in the East carry a smaller penalty than mothers raised in the West, whichever side of the old line they now live on (HC-05, citing an NBER working paper).

Quentin Lippmann, Alexandre Georgieff and Claudia Senik went inside the household. Their data is the Socio-Economic Panel from 1991 to 2012, and their question is what a couple does in the year after the wife's income crosses half the household total (Lippmann, Georgieff and Senik, *The Economic Journal*, 2020, per SY-06 and HC-06). The Preface told you the answer: in the West she does more housework, is more likely to leave

her job, and the couple is more likely to divorce, and in the East none of it. What I want to add is the size of the thing being tested. In 1992, two years after reunification, almost 30 per cent of East German wives out-earned their husbands, against 16 per cent in the West, and the gap has not closed (HC-06). The "doing gender" reaction, which Kan and Gershuny review across the western literature in the European Sociological Review in 2016 (HC-04), is not a human universal. It is a western one, in the German sense of western, and forty-one years of a different tax office was enough to switch it off.

Two more records, each a different instrument. Stefan Bauernschuster and Helmut Rainer put attitude questions to both halves of the country and found East Germans 22 percentage points more likely to hold egalitarian views on sex roles (Bauernschuster and Rainer, Journal of Population Economics, 2012, per SY-06). And the register of births. In the West in 1950 one child in ten was born outside marriage, 10.6 per cent, and by 1965 it was 5.8 (SY-06, citing the Federal Statistical Office through a secondary website). Across reunified Germany the figure in 2023 was 33.1 per cent, and in many eastern districts it is over half, over 60 in places (SY-06, same caveat). Sebastian Klüsener and Joshua Goldstein traced the divide back before the division and found the East ahead already, and the GDR then made the lead permanent (Klüsener and Goldstein, *Population, Space and Place*, 2016, per SY-06).

Put the records together and they say one thing clearly. Give a population a tax code, a timetable and a word for forty-one years, and you move its mothers' hours by a month and a half a year, its wives' willingness to out-earn their husbands by half, its attitudes by 22 points and its birth register by a factor of five, and the movement outlives the regime by a generation and travels in a suitcase. That is the culture layer, measured. It is the largest effect in this book.

Where the wall stops

Now the other column.

The fathers did not move. In the records that found 350 hours between the mothers, there is no gap between the fathers at all (SY-06). The East had abolished the breadwinner norm, pushed nine women in ten into

paid work, and left the men's hours exactly where the West had left them. Rosenfeld, Trappe and Gornick described what reunification had produced in both halves as a convergence not on the dual-earner household but on a modernised breadwinner: he full-time, she part-time (HC-06, SY-06). The East had been told, for forty-one years, what a mother was for. Nobody had told it what a father was for, and so a father went on being what he had been in 1949.

The kitchen did not move. HC-06 is blunt about the asymmetry: a state can move public behaviour, hours in the shift and the office, and meets what the document calls almost insurmountable friction at the private negotiation between a man and a woman in a flat. The East's double burden was the proof, and reunified Germany's family panel repeats it in both halves: the laundry and the cooking still mostly hers, the outdoor jobs mostly his, and the symmetrical couple reverting to the old split at the first birth almost invariably (chapter 7 has the FReDA figures and I am not going to repeat them). Whether the eastern version of equality raised women's satisfaction or merely moved their tiredness from the home to the factory is a question SY-06 lists as open, under the dry heading of the productivity paradox. I do not know. Neither does the document.

The young did not move, or stopped moving. Daniel Lois took nine waves of the German General Social Survey, 1982 to 2016, 26,389 people, and separated the effect of age from that of the year from that of the birth cohort (Lois, *Comparative Population Studies*, 2020, per SY-06). In the West the drift towards egalitarian attitudes stopped for men born around 1956 and later and for women born after 1966. In the East the younger cohorts also soften a little, from a higher level. I was born in 1975. The men of my cohort, in the West, are the first in forty years who are no more egalitarian than their fathers, and I say that about my own side of the argument before anyone else can.

And here are the two readings of the stop, and the chapter has to carry both. The first is Catherine Hakim's, from 2000: that once a society gives women contraception, equal law and white-collar work, they sort themselves by preference into roughly 20 per cent who put the home first, 20 per cent who put work first and 60 per cent who adapt and prefer part-time (SY-06). On that reading the West's part-time mother is not a victim of a tax code. She is the 60 per cent, and the East was a forty-one-year

interruption to it. The second reading is Corinna Frodermann's and her colleagues', from a survey experiment in 2020 called "Money Matters!": show respondents a vignette in which a mother's wage is what keeps the household afloat and the belief that mothers must be at home collapses on the spot (Frodermann et al., *Social Forces*, 2020, per SY-06). Preferences, on that reading, are what the price list looks like from inside.

I think the records settle the first half of this and not the second. A preference that moves 350 hours when you cross a border, keeps moving when you cross back, and spreads to the next desk at the office is not a preference in Hakim's sense. It is a price. But a father whose hours do not move in either direction, under either regime, is not a price either, or not one anybody has yet tried to change, and the years since Germany introduced an earnings-related parental benefit with months reserved for fathers, the *Elterngeld* of 2007, are too few for the records to say whether that was the missing crane (SY-06, which lists it as open).

So here is the tool, and it is a reading instrument rather than a lever. The wall shows you how to separate three things that every argument about men and women runs together. What moved under the regime is culture, and it is large, and it is the mothers. What did not move under either regime is either the body or a wall nobody has yet tried to take down, and the fathers and the kitchen are on that side of the line, and the records cannot yet tell which. And what the regime was never asked about, the accent in the room when the hours are finally equal, is the question the next three chapters and the last one take up. Measurable, statistical, ours. Now the case against, and it comes from both shelves at once.

THE STRONGEST CASE AGAINST

The chapter has argued that culture is the largest force in the book on the mothers' side, and that the fathers' stillness is either the body or an untried wall. The case against comes from both ends of the bookshop.

From the evolutionary shelf, the kibbutz. Lionel Tiger and Joseph Shepher went through three generations of occupational records on the Israeli kibbutzim, where childcare, laundry and cooking were collectivised and children slept in children's houses, and found that by the second and third generations, born into equality, about 75 per cent of women

had gravitated to education, health and services and about 75 per cent of men to agriculture, industry and management (Tiger and Shepher, *Women in the Kibbutz*, 1975, per HC-06). Melford Spiro, in 1979, found the same. And the pressure to bring the children back to the family flat to sleep came from the mothers, not from the men (HC-06). The sample began at 30 people, and the critics, Eva Rosenfeld in 1958 and van Teeffelen in 1977, point out that the kibbutz sat inside a patriarchal region and that women in heavy industry were tokens who were edged out (HC-06). But the shape is the opposite of the East German shape. Remove the price and the division comes back. Hakim's 60 per cent, on that reading, were the sabra generation, and the East's 89 per cent were a wage bill.

From the same shelf, the paradox. Armin Falk and Johannes Hermle measured economic preferences, from risk to trust, in 80,000 people across 76 countries and found the gap between the sexes largest in the richest and most equal places, correlating at 0.67 with the log of national income and 0.56 with their equality index (Falk and Hermle, *Science*, 2018, per HC-07). Chapter 5 carries the fight over the explanation, Fors Connolly included (HC-07). But if the pattern holds even in part, the German drift toward a modernised breadwinner after 1990 is not only the West's tax code winning. It is what freedom looks like, and the East's dual-earner household was the thing that needed a state to hold it up.

From the constructivist shelf, the harder blow, because it is aimed at the half of the chapter I like. The fathers did not move, I said, so perhaps the body is in the way. The opponent's answer is that nobody pushed them. The state did not challenge male privilege at home (HC-05), and no institution required men's domestic hours to rise (HC-06). A still father in a regime designed around mothers is not evidence of an accent. It is evidence of a regime designed around mothers. And the clinching comparison is Japan: 224 minutes a day of unpaid work for a woman against 41 for a man, the widest gap in the OECD, in a country where men put in 452 minutes a day of paid work, the longest (OECD 2020 figures per HC-06, sourced to a secondary column). Where men are not in the house, there is no private negotiation to lose. The kitchen does not stay with the woman because of her body. It stays with her because he is at the office, and the office is a tax code too.

I cannot refute either shelf from the German records alone. The kibbutz says the division returns when the price is removed. The East says

it stays removed when the price is removed and the regime is gone. They disagree. That is not a reason to pick one. It is the subject.

IN PRACTICE

The scale here is the state, and most of you do not run one, so let me speak to what a person does inside one, and then to what a state does if it is listening.

If your country splits a couple's income for tax, then the arithmetic is already in your kitchen whether or not you have read it, and the first thing to do is read it: work out what the second earner's first hour is actually worth after the schedule has taken his rate off it, and decide with that number on the table rather than with the gross. The failure mode is reading the number and treating it as nature. It is a line in a statute, it was passed in 1958 against other options, and Scandinavia chose the other options (HC-04).

If the school day ends at lunch and the crèche is a rumour, then one of you is going to be at the gate at a quarter to one, and in the German records it is her, and the records say the reason is the price list, not the person. Decide who stands there before the first birth and not after it, because the family panel says the symmetrical couple reverts at the birth almost invariably and does not revert back. The failure mode is the plan that assumes the timetable is negotiable. It is not. Only the person at the gate is.

If you are thinking of moving, then the records say the move changes the mother and not the father. A western mother who goes east tends to work like an eastern one within a few years, and tends to keep doing so if she returns. The counter-evidence is that the records can see hours and not happiness.

If you are a man born in the West around 1975, then the survey data say you are, on average, no more egalitarian than your father, and that the drift stopped around the year you were born. The failure mode is thinking it stopped because the argument was won.

If you run a firm in a breadwinner country, then your office is a border. Western mothers in firms where the share of eastern colleagues rose went back to work sooner after a birth, which means that the composition of the team is a family policy whether or not you meant it as one. The failure

mode is treating this as a lever: the document gives the direction and not the size, and an engineered peer effect is a different thing from one that happens.

Now the state, because the German records are the clearest instructions a state has ever been given.

If you want mothers in paid work, then the records say three things move them: the price of her hour, the length of the school day and the word for her, and the word alone accounts for most of the gap. A tax reform without the crèche tends to move little, a crèche without the word tends to move something, and the East moved everything, with methods no democracy should want. The failure mode is believing that mindsets change on their own once the law is equal. Lois's cohort data say they stall.

If you want fathers in the house, then no German regime has yet shown you how, and the honest instruction is that nobody has tried with the force that was used on mothers. The *Elterngeld* with its reserved months dates from 2007 and the records cannot yet say what it has done (SY-06). The counter-evidence is the Japanese column: where men's paid hours are the longest in the OECD, no family policy reaches the kitchen because the man is not in it. A father's hours at home are set at the office, and a state that wants to move them has to move the office first.

If you want the private sphere to change at all, then the records say the state's reach stops at the front door in both regimes, and the kibbutz says it stops there even when the front door has been removed. What moved the East's kitchens, to the extent they moved, was not a decree. It was that she out-earned him and nobody reached for the ironing board. Give couples that fact, through the tax code, and the rest is theirs, which is the only place it was ever going to be decided.

And if you are tempted to run the experiment again, in either direction, then remember what it cost. The East's equality was a labour policy with a wall around it. The West's family was a tax advantage with a word attached. Neither was chosen by the women it was built for, and the one thing the records are unambiguous about is that the women would have chosen differently from both, given, for once, the whole question.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- Ehegattensplitting as a price on the second earner's hour, from 1958, Steiner and Wrohlich 2004 (SY-06).
- The half-day school and the absent crèche as infrastructure that requires a parent at home (SY-06).
- Boelmann, Raute and Schönberg 2025: the collective model, the wage decomposition, the 73 per cent, the movers, the returners, the firm peer effect (SY-06).
- Lippmann, Georgieff and Senik 2020: the doing-gender reaction in the West and its absence in the East, 30 against 16 per cent (SY-06, HC-06).
- Kleven et al. 2019 and Kleven 2024: the German penalty in world context and by region of socialisation (HC-05).
- Fathers' hours unmoved in both regimes and the East's double burden (SY-06, HC-06, HC-05).
- Lois 2020: the cohort stall (SY-06). Lewis 1992: Germany as a strong male-breadwinner state by design (HC-04).

Decoration

- The school gate, the tax form, the two nouns as a symmetry.
- Birmingham and Essex (HC-04).
- The meeting of the two movements in the early 1990s (SY-06, weakly sourced).
- The pension line (chapter 7's figure).

Contested

- Whether the kitchen's stillness is the body or an untried wall (SY-06's open question, HC-05, HC-06, against the kibbutz reading).
- Hakim 2000 against Frodermann 2020: part-time as preference or price (SY-06).
- The kibbutz reversion as nature (Tiger and Shepher, Spiro) or as tokenism and context (Rosenfeld, van Teeffelen, Kanter) (HC-06).
- Kleven's German penalty, 61 (HC-05) against 60 (FA-05).
- The Rabenmutter and Schmarotzer definitions, sourced to a forum thread (SY-06).

NUMBERS TO CARRY

1. About 1,800 hours a year of paid work for an East German mother, 350 fewer for a West German one, no gap for fathers, in a 50 per cent sample of social-security records. Equalising wages would nearly double the gap. Stigma accounts for 73 per cent of it (Boelmann, Raute and Schönberg, *AEJ: Applied Economics*, 2025, per SY-06).

2. Wives out-earning husbands in 1992: almost 30 per cent in the East, 16 in the West. In the West, crossing half the household income raises her housework, her exit and the divorce risk the next year, in the East none of these. SOEP 1991 to 2012 (Lippmann, Georgieff and Senik, *The Economic Journal*, 2020, per HC-06 and SY-06).
3. Long-run child penalty about 61 per cent in West Germany (60 per FA-05), about 20 in the United States, 21 to 27 in Scandinavia (Kleven, Landais, Posch, Steinhauer and Zweimüller, *AEA Papers and Proceedings*, 2019, per HC-05).
4. East Germans 22 percentage points more likely to hold egalitarian sex-role attitudes (Bauernschuster and Rainer, Journal of Population Economics, 2012, per SY-06). The egalitarian drift in the West stalled for men born around 1956 and women born after 1966, ALLBUS 1982 to 2016, 26,389 respondents (Lois, *Comparative Population Studies*, 2020, per SY-06).
5. Births outside marriage: 10.6 per cent in the West in 1950, 5.8 in 1965, 33.1 nationally in 2023, over 50 to 60 in many eastern districts (Destatis via SY-06, and Klüsener and Goldstein, *Population, Space and Place*, 2016).

THE LIBRARY

- SY-06 Germany: two gender models as a natural experiment. The splitting, the timetable, the two nouns, Muttipolitik, Boelmann, Lippmann, Klüsener, Lois, Hakim against Frodermann, the Elterngeld question.
- HC-05 Feminisms of sameness and difference, the German story. The Hausfrauenehe, Paragraph 218, the Stern action, the 1974 law and the 1975 court, the 1972 GDR law, the 1993 compromise, Kleven 2019 and 2024.
- HC-06 Natural experiments. The 89 and 56 per cent, Lippmann's 30 against 16, Trappe and Rosenfeld, the kibbutz, Japan's 224 against 41.
- HC-04 Industrialisation and the separation of home and work. Davidoff and Hall, the family wage dispute (not used), Lewis's typology, Kan and Gershuny.
- HC-07 The gender-equality paradox. Falk and Hermle 2018 and Fors Connolly 2020, for the case against only.

CHAPTER 14

Queens at War

A woman is standing up in a village council meeting in West Bengal, and the room is not used to it. The post she holds, the chair of the council, was not hers to win in the ordinary way. In 1993 India amended its constitution so that one chair in three at the head of a village council would be reserved for a woman, and which villages got a reserved chair was decided at random, so that nobody could later say that the places with women in charge were the places that had wanted them. She is reading out the budget. More of it than last year goes on drinking water, which is what the women of this village complain about most when anyone asks them. Along the wall sit those women, who in an ordinary year get through the whole meeting without a word, and this year they are speaking up about twice as often as they do when a man has the chair. Nobody instructed them to. The chair changed, and the room changed with it.

Hold that picture, and a second one beside it, four centuries older and a continent away. Somewhere in the 193 European reigns that two economists would one day sort into a spreadsheet, a queen is at war. Not defending. Attacking. Her husband is running the treasury and the army while she runs the state, and the two of them together are a more dangerous neighbour than any king whose consort is kept out of the room. Across all 193 reigns, from 1480 to 1913, the chance that a polity is at war in a given year is 39 percentage points higher under a queen than under a king.

Both pictures are true, and both come from the same kind of evidence: the natural experiment, the lottery, the accident of birth order, the thing

that decides who holds power without asking anyone's opinion of women. That is the only kind of evidence this chapter trusts, because the countries which put women in charge are already different from the countries which do not.

If you came here hoping the chapter would settle whether women should run states, it does something less satisfying and more useful. It shows what changes when they do, which is a great deal, and what does not, which is the willingness to fight. The queen is the chapter's standing image because she breaks the folk theory on both shelves at once. She is not the gentler sex in a crown. She is also not a man in a dress. She is a person with a different set of constraints, and the constraints, it turns out, are most of the story.

So: what states do when women hold power, where the law still keeps two columns, and why the best predictor of a state's peace is how it treats its women, which is not, and this is the whole point, the same as saying that women are peaceful.

The lottery

The reason the Indian village councils matter so much is the lottery. Raghabendra Chattopadhyay and Esther Duflo, economists, took 265 village councils in West Bengal and Rajasthan after the 1993 reservation came in and compared the ones that had drawn a woman chair with the ones that had not (*Econometrica*, 2004). Because the draw was random, any difference in what the councils did could be laid at the door of the chair's sex and nothing else. In West Bengal, where women's complaints ran to drinking water, the councils with women in the chair spent more on water. In Rajasthan, where the men travelled for work and wanted roads, the women's councils spent less on roads, though the two reports behind this chapter disagree about the roads and I carry only the water. And the women of the village spoke at the meetings about twice as often. That is the finding I would keep if I could keep only one, because it is about the room, not the chair.

What happens when the lottery ends? Rikhil Bhavnani looked at Mumbai wards where a reserved term had lapsed and found that a woman was about five times as likely to win the next open election there as

in a ward that had never been reserved (*American Political Science Review*, 2009). The quota did not just put a woman in office for a term. It let the voters find out what a woman in office looks like, and the parties that she can win.

Now scale up. Sonia Bhalotra and Irma Clots-Figueras took the Indian state legislatures and asked what a larger share of women in the assembly does to the children of the state (*American Economic Journal: Applied Economics*, 2014). A ten-point rise in women's share of the seats cut neonatal mortality by 2.1 percentage points, a third of the baseline, through more antenatal visits, more iron supplements and more births in a facility. Then Bhalotra, with Damian Clarke, Joseph Gomes and Atheendar Venkataramani, went global, using the staggered dates on which developing countries adopted parliamentary quotas for women (*Journal of the European Economic Association*, 2023). After a quota, the maternal mortality ratio fell by 8 to 12 per cent, skilled attendance at birth rose by 5 to 8 points, fertility fell by 6 to 7 per cent, and young women stayed in school half a year longer. A companion analysis looked back at the United States in 1937, when the first antibiotics, the sulfa drugs, arrived, and found that maternal deaths fell 15 per cent further in the states that had given women the vote early than in the states that had given it late.

So the first finding is in the culture layer, and it is as solid as anything in this book. Change who holds the chair and the budget moves towards water, clinics and children. It is a finding about priorities, not about virtue, and the difference matters in a moment.

One more from the Indian file, because it answers the obvious objection. If women spend on health and schools, do they neglect the economy? Thushyanthan Baskaran, Sonia Bhalotra, Brian Min and Yogesh Uppal took 4,265 close elections to Indian state assemblies, the ones a woman narrowly won against a man or narrowly lost, so that the winner's sex was as good as a coin toss (*Journal of Economic Growth*, 2018 and 2023). Constituencies that got a woman grew 1.8 percentage points a year faster, measured by the light visible from satellites at night, the economist's way of counting without asking. The men in the same close races were three times as likely to be carrying criminal charges, and while in office they accumulated assets about ten points a year faster than the women did.

That last number sounds like a verdict. It is not one, and the next section is about why.

The fairer sex, audited

Here is the finding that launched a thousand columns. David Dollar, Raymond Fisman and Roberta Gatti put the share of women in a country's parliament next to its perceived corruption across more than a hundred countries and found that a one-standard-deviation rise in women's share went with a fifth of a standard deviation less corruption (2001). The paper's title asked whether women really were the fairer sex, and a great many people answered yes before reading it.

The answer that survived is more interesting. Justin Esarey and Gina Chirillo noticed that the relationship held in democracies and vanished in autocracies (*Politics and Gender*, 2013, under the title "Fairer Sex or Purity Myth?"). If women were less corrupt by nature, the regime type should not matter. Esarey and Leslie Schwindt-Bayer then showed, across 76 democratic-leaning countries, that women reduce corruption only where the odds of getting caught and punished at the ballot box are high (*American Journal of Political Science*, 2018). Two overlapping mechanisms do the work. One is statistical: women on average take fewer risks, and corruption in an accountable democracy is a risk. The other is cultural, and it is the sharper of the two. Tiffany Barnes and Emily Beaulieu found in experiments that voters hold women to a higher ethical standard, so a woman caught in a scandal is punished harder at the polls than a man caught in the same one (2014 and 2018). Women in politics are cleaner, where they are cleaner, partly because they are watched more closely. That is a fact about the audience, not the molecule.

And the risk aversion that does half the explaining is itself a smaller thing than the folklore. Rachel Croson and Uri Gneezy's review of the laboratory games found women more averse to risk and more sensitive to context on average (*Journal of Economic Literature*, 2009). But Julie Nelson went back through 147 estimates from 25 studies of risk aversion and found that the studies showing a difference had been published and cited, the ones showing none had been quietly filed, and once you correct for which papers made it into print the gap shrinks a good deal (*Feminist Economics*, 2015). The variation among women, and among men, dwarfs the variation between them. Then the context-sensitivity itself was put to a large replication by Stefano DellaVigna and Devin Pope, who found that 93 per cent of the between-subject comparisons in their data contradicted

it (2022). Each of these is a behaviour-layer finding. Each has an overlap you could drive a bus through.

There is one more filter between the population and the podium, and it is the one the folk theory always forgets. The women who reach the top of national politics have come through the same selection machine as the men. One of the reports behind this chapter puts it drily: female heads of state frequently sit in the tails of the distributions, and are statistically hard to tell from their male colleagues on competitiveness, risk tolerance and the appetite for dominance. The central bankers in chapter 10 were the clearest case, with women on monetary committees voting for *higher* rates than men at the same expected inflation, which the authors read as selection or as signalling, and in neither case as temperament. An average is a fact about a population. The woman at the dispatch box is one person, and she got there by not being average.

Which brings us to the queens.

Queens

Oeindrila Dube and S. P. Harish, economists, wanted a case where women held supreme power and nobody had voted for them, so that the countries which choose women could not be confused with the women they choose. Hereditary monarchy is that case. They assembled 193 reigns across 18 European polities between 1480 and 1913 and asked a plain question: is a polity more or less likely to be at war in a given year when a queen holds the throne (*Journal of Political Economy*, 2020, with the one-word title "Queens")? Even here selection could creep in, since a queen tends to inherit when there is no brother, which might mark an unusual family or an unusual moment. So they used two instruments that nobody chooses, whether the previous monarch had a first-born son and whether the previous monarch had a sister, to isolate the part of a queen's accession that is pure accident of birth.

The answer was 39 percentage points. A polity under a queen was that much more likely to be at war in any given year than a polity under a king. Then the finding splits in two, and the split is where the chapter earns its keep. Unmarried queens were more likely to be *attacked*: their neighbours seem to have read a woman alone on a throne as an opening.

Married queens were more likely to be the aggressors, and more so than married kings. Dube and Harish's explanation is a division of labour. A married queen could put her husband in charge of the army and the exchequer and keep the state for herself, in effect running a dual executive, which gave her both more capacity for war and more allies to wage it with. A married king, in the same centuries, did not hand his wife the treasury. The queen was not more warlike as a woman. She was more thoroughly staffed.

That is a culture-layer finding dressed as a biology-layer surprise, and I think it is the most useful single result in this chapter, because of what it does to both shelves. To the shelf that says women in power would end war, it says: here are four centuries of women in power, and they fought more. To the shelf that says women cannot or do not wage war, it says the same thing. And to this book it says that the constraint is the variable. Give a leader a capable co-executive and the state fights. Take the constraint away and the sex of the person in the chair tells you very little about whether the cannon is loaded.

Set that beside the correlational finding that has become the slogan of a movement. Valerie Hudson and her colleagues built a database of women's lives, the WomanStats project, 130,000 data points across 175 countries, and reported that the physical security of a state's women predicts that state's peacefulness better than its wealth, its democracy or its religion (*Sex and World Peace*, 2012). States that tolerate violence against women fight more, inside and out. The critics' objection is the one you would expect, and the report behind this section gives it fairly: rich, stable, democratic states both protect their women and avoid wars, so the arrow could point either way, or both ways from a third thing. I take the objection seriously and still think the finding has a shape worth carrying, because of what Dube and Harish add to it. If women in power are not the peaceful sex, then the Hudson correlation cannot be about the temperament of women. It has to be about what a state is like when it treats half its people as property. That is a claim about institutions, and institutions can be changed.

What kind of state is that? Here the deep-history databases, the ones chapter 2 used for the ledger of who cooks, give an answer about war that chapter 2 did not need. Across the Standard Cross-Cultural Sample, which George Murdock and Douglas White assembled in 1969, there

is not one society in which women hold the military leadership. Peggy Reeves Sanday, in 1981, scaled 156 of those societies by women's economic control, political voice and symbolic standing and found a pattern she called inner and outer orientation (*Female Power and Male Dominance*). Societies that lived off gathered plants and had long stretches of calm tended to share power. Societies under stress, in famine, on the move or at endemic war, came to rely on the two things male bodies do best, hunting large animals and fighting, and the men who did those things took the rest. In Sanday's data that outer orientation travels with high rates of violence and with a culture in which rape is common. William Divale and Marvin Harris had already given the pattern a name in 1976, the male supremacy complex: a society that needs young men to die for it tends to pay them in patrilocal residence, extra wives, a monopoly on politics and the subordination of its women (*American Anthropologist*). Bobbi Low, with controls for shared ancestry, found that pathogen load and male deaths in war were the two best predictors of a society practising non-sororal polygyny. And the arrow can run the other way. Joseph Henrich, Robert Boyd and Peter Richerson argued that although about 85 per cent of the societies in the ethnographic record permitted a man more than one wife, normative monogamy spread because of what it did to the group: fewer unmarried men without prospects, less murder, rape and robbery, a smaller age gap between spouses, and fathers who invested in children rather than in the next wife (*Philosophical Transactions of the Royal Society B*, 2012).

Which is the point at which this chapter should turn to sex ratios, surplus men and the stability of states, and cannot.

The section on sex ratios, polygyny and social stability is not yet written. Its research report is missing from the library.

Whatever the gap turns out to hold, the rest of the evidence does not run through female gentleness. Martin King Whyte took 93 pre-industrial societies in 1978 and tested the assumption underneath every "status of women" index ever built, that a society which gives women one kind of standing gives them the others (*The Status of Women in Preindustrial Societies*). It does not. His factor analysis found ten dimensions of women's standing that barely moved together, property, domestic authority, ritual inclusion, sexual freedom, political voice and the rest. A society could give women the home and the temple and keep them out of the council and the war band entirely. There is no single thing called the position of wo-

men that rises and falls as one. There are ten things, and a state can be generous on seven and murderous on three.

Two columns in the statute book

Now turn the question around. Not what women do to the state, but what the state still does to the sexes, after seventy years of laws that promise to treat them alike. The law has two columns, and some of them are written in bone.

Start with the one that cannot be legislated away. Norway, in 2013, passed a law conscripting women as well as men, the first member of NATO to do so, and it took effect in 2015. The armed forces put men and women in the same dormitories, and Nina Hellum of the Norwegian Defence Research Establishment went to the camps at Skjold and Heggelia to see what happened (2014). Shared rooms produced something she called de-genderisation, a sibling-like team feeling and less harassment, while the all-female rooms reported more drama. That is a culture-layer result and a cheerful one. Then the physiology layer, which does not read Norwegian statutes. By 2020 women were 14 per cent of the force against a target of 20. And in Israel, where the Anti-Aircraft Forces put men and women through identical basic training, Aharon Finestone, Charles Milgrom, Daniel Moran and colleagues counted the stress fractures: 23.9 per cent of the women, 33 recruits, against 11.2 per cent of the men, 42 recruits, a relative risk of 2.13, and the women's fractures sat in the tibia, the femur and the pelvis, the bones an infantry soldier walks on (*Military Medicine*, 2005). In light infantry the same group later found 12.0 per cent against 1.3, a relative risk above nine (2014). Across integrated militaries the hazard in the first six months of training runs at more than four times. Lower bone density and smaller muscle mass are measurable. The law can make the duty identical. It cannot make the skeleton identical, and a state that pretends otherwise breaks its own soldiers. And still, as the Israeli data also show, individual women with high baseline fitness complete the tours, which is the overlap doing its usual work against the average.

The second column is written in culture, and it runs the other way. Sonja Starr took the United States federal criminal cases and followed each one from arrest to sentence, so that the leniency at every stage could be ad-

ded up rather than only the judge's last decision (*American Law and Economics Review*, 2014). Holding the arrest offence and the criminal history constant, men received sentences 63 per cent longer than women. Women were twice as likely to avoid prison altogether if convicted. About 30 per cent of the gap in incarceration was unexplained by anything in the file, and the whole gap was about six times the size of the racial gap found by the same kind of study. France, where every conviction from 2000 to 2003 was in the dataset, gives the finding its cleanest form: women's prison sentences were 15 days shorter on average, a third of the 45-day mean, and among pairs of co-offenders, one man and one woman, convicted of the same crime by the same judge on the same day, the man got the longer sentence in 22 per cent of cases and the woman in 4. One more standard deviation of women on the bench shrank the gap by a tenth. The paternalism is mostly the men's, and it is a behaviour-layer average, women commit far less violent crime, applied to one defendant, which is exactly what this book keeps saying not to do.

The family court keeps the same two columns. The documents give a figure of 80 to 84 per cent of contested custody cases ending in sole physical custody for the mother, with no primary source, and I carry it only to say what the law's default has been. Against that default the evidence has been pushing for twenty years. Robert Bauserman's meta-analysis found children in joint custody closer to children of intact families than those in sole custody, and found, against the standard objection, that joint-custody parents reported less conflict, not more, by 0.24 of a standard deviation (*Journal of Family Psychology*, 2002). Linda Nielsen's review of 60 studies found 34 in which joint-custody children did better on every measure, 14 in which they did as well or better, and 6 in which they did worse on one measure (2018). When Jennifer McIntosh's Australian studies of 2010 and 2013 led courts to deny overnight stays to fathers of infants, Richard Warshak published a consensus report signed by 110 researchers saying the evidence did not support the denial and that overnights kept fathers from dropping out (*Psychology, Public Policy, and Law*, 2014). Chapter 8 had the physiology of the father's brain. Here is the law catching up with it.

So when does a mixed body decide differently from a single-sex one? Chapter 10 found the laboratory answer unstable. The state-scale answer is steadier, and it comes with a mechanism from the animals. Larissa Con-

radt and Timothy Roper modelled group decisions in which members have different interests and showed that despotic decisions, one animal deciding for all, cost the group more than shared ones, because a shared decision lands nearer the middle of the group's interests and avoids the extremes (*Nature*, 2003). That is the whole case for a mixed body in one sentence: not that women are wiser but that two distributions averaged together have a shorter tail than one. The human evidence runs the same way where the stakes are highest. Astghik Mavisakalyan and Yashar Tarverdi found that more women in a parliament goes with lower carbon emissions (*European Journal of Political Economy*, 2019), and Supriya Garikipati and Uma Kambhampati, matching 194 countries in the first wave of the pandemic of 2020, found that the women-led ones locked down earlier and lost about half as many people (*Feminist Economics*, 2021). And in the Capital Jury Project, 1,200 interviews with jurors from 353 death-penalty cases in 14 American states, William Bowers and colleagues found that white male jurors were the readiest to sentence a defendant to death and that women on the jury weighed mitigating evidence more heavily, so that a mixed jury kills less often than a male one.

Here is the turn. In the nineteenth century men held every vote and every seat, and they gave the vote and the seats away. Matthias Doepke and Michèle Tertilt asked why, and their answer is in the title of their paper, "Women's Rights: What's in it for Men?" (*Econometrica*, 2009). As the return on education rose, a man's best bet for his own line was to have his children schooled, and women spend a larger share of household money on children than men do, so men voted to give women property rights, and in time the vote, to get the money where they wanted it. It is not a romantic account of emancipation. It is a true one, as far as it goes, and it hands you the tool for this chapter's scale. Nobody at the top of a state has ever shared power because they were asked nicely. They share it when it pays, or when a lottery makes them, or when a queen with a capable husband is on the border. If you want a state to behave differently towards its women, do not argue about what women are like. Change what it costs to treat them badly.

THE STRONGEST CASE AGAINST

The chapter has argued that putting women in power changes what states do, and that the sex of the person in the chair tells you little about war. The strongest case against runs through both halves, and it is a case about confounding.

Begin with the corporate boards, because they are the best-measured bodies on earth and the effect there is close to nothing. Corinne Post and Kris Byron pooled 140 studies of women on boards and found a small positive relation with accounting returns, strongest where shareholders are well protected, and a relation with market performance near zero (*Academy of Management Journal*, 2015). Vathunyoo Sila, Angelica Gonzalez and Jens Hagendorff found no effect of women directors on the equity risk of a large sample of American firms at all (*Journal of Corporate Finance*, 2016). David Matsa and Amalia Miller's Norwegian quota study, the most-cited sign of a "female style", found boards with more women laid off fewer workers and earned less in the short run (*American Economic Journal: Applied Economics*, 2013), which is a priority, not a competence, and it is the only robust board finding the documents offer. If sex composition does so little where the data are best, the burden is on the political findings to show they are not the same thing seen through a thinner lens.

And the political findings have an identification problem that the Indian lotteries solve and nothing else does. Leah Windsor and colleagues made the point about the pandemic and the climate: the countries that elect women are high-trust, low-hierarchy countries, and those are the countries that lock down early and tax carbon, whoever is in the chair (2020). The Hudson correlation is open to the same reading, and the chapter's own sources say so. Every instrument in the climate papers is an argument, not a lottery.

Then the behaviour layer itself may be thinner than the chapter let on. Nelson's correction for publication bias is not a footnote: it means the risk-aversion gap that half the corruption story rests on was partly manufactured by the journals (*Feminist Economics*, 2015). DellaVigna and Pope's 93 per cent is a replication failure for the idea that women's social preferences bend to context. And the one physiology-layer finding the documents offer in women's favour, Shelley Taylor's tend-and-befriend pattern from 200 studies and 14,548 participants (*Psychological Review*, 2000), rests on oxytocin, which the Preface has already reported does not do what it was sold as doing. Strip those out and what remains is: women

in office spend differently. That is enough for the chapter's first half. It is not evidence of a different nature.

Finally the queens could be read against me. Thirty-nine points is a vast effect from 193 reigns, and Dube and Harish's mechanism, the husband as co-executive, is an inference about staffing from the pattern of who attacked whom. A sceptic could say the finding is about early-modern Europe and tells us nothing about a cabinet. She would be right that it tells us nothing about women. It was never about them. It tells us about constraints, and if constraints are the story then the chapter's critics and I are closer than either of us would like.

IN PRACTICE

This chapter's scale is the state, which most readers do not run but all of them vote in, pay for, are sentenced or conscripted by, and sit on the juries and councils of. The conditionals are for that reader, and for the few who draft the rules.

If you are deciding whether a quota is worth the cost, then the Indian evidence says the gain is not mainly the term the quota buys but the term after, when a woman is five times as likely to win the open seat. The failure mode is a quota without an exit: Bhavnani's effect was measured where reservations lapsed, and a seat that is reserved forever teaches voters nothing. The counter-evidence is the token seat: a quota that puts a faction's women in the chair tends to buy the faction, not the change.

If the argument for women in office in your country is that they are more honest, then stop making it, because it is true only where the press is free and the next election is close, and it tends to collapse the moment the accountability does. Esarey and Schwindt-Bayer's 76 countries are the counter-evidence. The failure mode is the one Barnes and Beaulieu measured: the honest-woman story raises the price of her first scandal, so the argument that put her in tends to be the argument that throws her out.

If you are reading a correlation between women's rights and peace, then read it as a fact about the state that violates its women, not as a promise about the women it would elect. The queens are the counter-evidence, 39 points of it. The failure mode is the peace movement that puts a woman in charge of the war ministry and is surprised.

If your state writes the same physical standard for both sexes in a trade that loads the skeleton, then expect the Israeli numbers, roughly double the fractures in mixed training and nine times in light infantry, and build the training to the bone rather than to the statute. The failure mode on one side is a standard lowered for women, which breaks the women it admits and the trust of the men beside them. The failure mode on the other is a standard pretended to be sex-blind, which breaks women in the first six months and calls it attrition. The counter-evidence is the individual woman who passes the light-infantry tour, and she exists in the same Israeli data, which is why the standard should be a standard and not a sex.

If your courts are handing men sentences 63 per cent longer for the same file, then that is the behaviour-layer average, women commit less violent crime, applied to one man in a dock, and the cure is the same as everywhere in this book: sentence the person, not the distribution. The failure mode is the correction that over-shoots and discovers, as the economic-marginalisation theorists predict, that equality for women in the dock means more women in prison, which few of the people asking for it had in mind. The counter-evidence is the French bench: more women judges narrowed the gap, which suggests the leniency is a paternalism that a mixed body tends to dilute.

If you sit on a body that decides something irreversible, a jury, a board, a committee that goes to war, then the case for mixing it is Conradt and Roper's, not Hudson's: two distributions averaged have a shorter tail than one. The failure mode is the token, the single woman in a room of twelve, who in the jury data and in chapter 10's teams tends to speak less and change nothing. The counter-evidence is the central bank, where the women were the hawks, so a mixed body tends to be less extreme in ways you did not predict and may not want.

If you want power shared at the top of a state, then ask what it costs the people holding it not to share, because that, in Doepke and Tertilt's account, is the only reason it has ever happened. The failure mode is the argument from virtue. It has a four-hundred-year record, and the queens are in it.

And if you were hoping this chapter would tell you that a world run by women would be a gentler one, then I don't know, and the best evidence I have says that a world in which it is safe to be a woman tends to be a gentler one, which is a different sentence, and the only one I can sign.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing

- The Indian lotteries: Chattopadhyay and Duflo 2004, Bhavnani 2009, Bhalotra and Clots-Figueras 2014, Baskaran et al. 2018/2023 (SO-01, SO-07).
- Quotas and maternal mortality: Bhalotra, Clarke, Gomes and Venkataramani 2023 (SO-01).
- Corruption as accountability, not purity: Esarey and Chirillo 2013, Esarey and Schwindt-Bayer 2018, Barnes and Beaulieu (SO-01).
- Queens: Dube and Harish 2020, with the married/unmarried split (SO-01).
- Whyte 1978's ten dimensions and Sanday 1981's outer orientation (DH-07).
- Starr 2014 and the French co-offender pairs (SO-04).
- Finestone 2005 and 2014 (SO-04).
- Conradt and Roper 2003 as the mechanism for mixed bodies (SO-07).
- Doepke and Tertilt 2009 as the turn to the present (SO-05).

Decoration

- The two opening scenes, the woman in West Bengal and the queen with her husband at the exchequer (the findings carry, the scenes are mine).
- The night-light satellites.
- The Norwegian dormitories.

Contested

- Whether the Hudson correlation is cause, consequence or both from a third thing (SO-01's own critics, SY-05 for the figures).
- The roads finding in Chattopadhyay and Duflo (SO-01 against SO-07).
- Whether the climate and pandemic results survive the culture-that-elects-women objection (Windsor 2020).
- Whether the risk-aversion gap survives Nelson 2015, and whether tend-and-befriend (Taylor 2000) survives the Preface's oxytocin verdict.
- The 80 to 84 per cent custody figure (no primary source) and the French study's attribution ("Breda 2020", table only), both SO-04.
- Everything that belongs in the missing SO-02 report.

NUMBERS TO CARRY

1. 39 percentage points: how much likelier a polity was to be at war in a given year under a queen than under a king, 193 reigns, 18 European polities, 1480 to 1913 (Dube and Harish, *Journal of Political Economy*, 2020, per SO-01).

2. 2.1 percentage points, a third of the baseline: the fall in neonatal mortality for a ten-point rise in women's share of an Indian state legislature (Bhalotra and Clots-Figueras 2014, per SO-01). 8 to 12 per cent: the fall in maternal mortality after a national quota (Bhalotra et al. 2023, per SO-01).
3. Twice as often: how much more women spoke in village council meetings with a woman in the chair, 265 councils (Chattopadhyay and Duflo 2004, per SO-01 and SO-07). Five times: a woman's odds of winning a seat once reserved against one never reserved (Bhavnani 2009, per SO-01).
4. 63 per cent longer: men's federal sentences for the same arrest offence and history, and twice the odds of avoiding prison for women (Starr, *American Law and Economics Review*, 2014, per SO-04). 22 against 4 per cent: the share of French mixed co-offender pairs in which the man, or the woman, got the longer sentence (SO-04, attribution open).
5. 23.9 against 11.2 per cent, relative risk 2.13: stress fractures in women and men under identical Israeli anti-aircraft training (Finestone et al., *Military Medicine*, 2005, per SO-04). 12.0 against 1.3, relative risk 9.23, in light infantry (Finestone et al. 2014, per SO-04).

THE LIBRARY

- SO-01 Women in political power and what states do: the lotteries, the queens, corruption.
- SO-04 Where the law treats the sexes differently: conscription, sentencing, custody, leave.
- SO-07 Mixed versus single-sex decision bodies and collective risk: Conradt and Roper, climate, pandemic, juries, and the board nulls for the case against.
- DH-07 Cross-cultural databases: Whyte, Sanday, Divale and Harris, Low.
- SO-05 Cultural evolution: Henrich, Boyd and Richerson on monogamy, Doepke and Tertilt on men yielding rights (state and world scale only).
- SO-02 Sex ratios, polygyny and social stability: missing.

CHAPTER 15

The Long Argument

A girl of fifteen sits at a desk in a school in Finland, in the spring of 2018, with a booklet in front of her that hundreds of thousands of other fifteen-year-olds in dozens of countries are filling in the same year. The mathematics is done. Now come the questions about herself. Whether she enjoys reading. Whether, when she does badly at something, it is because she lacks the talent for it. She ticks a box. The booklet goes into a pile, the pile into a crate, the crate into a database with a four-letter name, and from there her tick travels to Paris, to be laid beside the ticks of girls in Tunis and Seoul and Lima and argued over by people she is never going to meet.

She does not know it, but she has just joined the oldest argument in this book.

It was being conducted in Greek in the fifth century before Christ, in a treatise on the diseases of women. It was conducted in Latin by an abbess on the Rhine, in Arabic by a physician in Damascus, in classical Chinese by a Tang doctor who decided that women needed their own medicine, and in Hebrew by rabbis who found that two categories were not enough for the bodies in front of them. It was conducted in law codes written on bamboo and in dowry contracts in Cairo and in the stone of a female pharaoh's temple, where she is carved with a beard. It is being conducted now in American journals, in the language of standard deviations, and it has not been settled, because the people having it have never quite agreed what it is about.

Someone, long ago, filed it. The file has been added to ever since, by

physicians and lawgivers, by rabbis and abbesses, by anthropologists and economists, and at some point in the last fifty years the whole thing was labelled reactionary and pushed to the back. The Preface promised that this book would open the drawer. The chapters so far have taken out one folder at a time. This chapter pulls the drawer all the way out and sets it on the table.

Pull up a chair. It is heavier than you expect, and older, and the oldest folder is at the bottom.

The oldest folder

The first people to write the argument down were doctors, and the first thing they wrote down was the obvious. The Hippocratic writers of the fifth and fourth centuries before Christ held that a woman's flesh was looser and spongier than a man's, that it took up more fluid, and that the monthly bleeding was the body's way of clearing the surplus (HC-03). That is a physiological claim, made from observation, and the half of it that says women menstruate and men do not has held up rather well.

The explanation did not. Aristotle, in the *Generation of Animals*, decided that only the male contributed seed, the female contributing matter, as a carpenter contributes shape to wood or rennet contributes curd to milk, and that a female was what happened when there was not enough heat in the womb to finish the job, an infertile male, a departure from the type (HC-03). Galen, in the second century of our era, gave the woman seed of her own, colder and weaker, and described her organs as the man's turned inward (HC-03). For most of two thousand years the medicine of Europe, Byzantium and the Islamic world argued inside the space those two men had marked out.

In 1990 the historian Thomas Laqueur, in *Making Sex*, told the story this way: that until the eighteenth century there was really only one sex in Western medicine, the male, with the female a lesser version on the same ladder, and that the modern idea of two distinct and opposite sexes was invented, not discovered, when Enlightenment talk of equality made the old hierarchy need a new foundation (HC-03). The book became a founding text for the view that sex itself is a cultural artefact. It was also, say the historians of medicine, wrong. Katharine Park and Robert Nye

answered it in 1991 under the title "Destiny is Anatomy", Joan Cadden in 1993, Michael Stolberg in 2003 in the journal *Isis*, and Helen King in 2013 in *The One-Sex Body on Trial* (HC-03). Stolberg's evidence is the plainest: anatomists had drawn a distinctly female skeleton, pelvis and all, by 1600, Felix Platter in 1583 and Caspar Bauhin in 1597, two centuries before the politics Laqueur said required it (HC-03). King's point is subtler and matters more for this book. Laqueur, obsessed with the genitals, had treated menstruation as a minor local case of general bodily flux, when to every pre-modern physician it was the defining fact of female physiology (HC-03). Laqueur replied in the same journal in 2003 and conceded little (HC-03). The settled view among historians is that the metaphors changed and the observation did not. A doctor can call the womb an inverted penis and still know perfectly well which patients bleed.

So what did the old doctors see that holds? Three things, in three traditions. In China, the Tang physician Sun Simiao, in the *Beiji Qianjin Yaofang* of about 651, set women's medicine apart as a discipline of its own, and his reason was not cosmological but physiological: menstruation, pregnancy, birth and recovery were stresses the male body never met and needed their own pharmacy (HC-03). The *Si Wu Tang* he and his successors prescribed for painful periods is still prescribed (HC-03). In Salerno, the collection known as the *Trotula* turned out, when Monica Green edited it in 2002, to be three texts, one of them the work of a real woman, Trota, whose treatments are practical, almost untouched by humoral theory, and who stated that infertility could lie in the man's seed as often as in the woman's womb (HC-03). And on the Rhine in the twelfth century Hildegard of Bingen, abbess and polymath, wrote in the *Physica* and the *Causae et Curae* of a woman's pleasure as part of conception, and then did something no one had done before her: she described a phlegmatic man and a sanguine woman, temperaments that crossed the sex line, and a varying vital sap she called *viriditas* that made each body a point on a range rather than a type (HC-03). An abbess, in the vocabulary available to her, had drawn the Overlapping Bells.

The Islamic physicians kept the argument moving. Ibn Sina, Avicenna, who lived from 980 to 1037, held in the *Canon* that there were two seeds and that a woman's pleasure was required to release hers, and borrowed Aristotle's rennet for the mixing (HC-03). Two seeds is closer to the truth than the ladder.

Two of their observations can be checked against modern numbers, and both pass, with a caveat each. The old texts noticed that women's moods and bodies shifted before bleeding. A systematic review by Direkvand-Moghadam and colleagues in 2014, in the *Journal of Clinical and Diagnostic Research*, pooled 17 studies and 18,803 women and found premenstrual syndrome in 47.8 per cent of them worldwide (HC-03). The caveat is the spread: 12 per cent in France, 98 per cent in Iran (HC-03). A condition of the physiological layer is being reported through a cultural filter so coarse that the filter is a finding in its own right. The old texts also noticed that a nursing mother did not conceive. A Cochrane review by Van der Wijden and colleagues in 2003 found the lactational amenorrhoea method more than 98 per cent effective in the first six months when the baby takes nothing but the breast, with pregnancy rates between 0.45 and 2.45 per cent across the trials, among them 236 Chilean women studied in 1988 and 419 Rwandan couples in 1996 (HC-03). The caveat is the condition: nothing but the breast, and only six months.

What they got wrong is also instructive. The woman as the cold sex, which the slightly higher resting metabolism of men turns on its head (HC-03). The wandering womb, which was supposed to roam the body and press on other organs. The womb does not wander. It has, as far as anyone can tell, never been anywhere.

The household ledger

Above the medical folder sits the legal one, and it is thicker, because every civilisation that could write wrote down who was allowed to own what. Here the layer is culture throughout, and the first thing the ledger shows is that the written rule and the lived arrangement are two different columns. Chapter 2 met Martin Whyte's finding that a society's treatment of women is not one thing but ten, which do not travel together. The civilisations fill in his table.

Athens gave a woman a guardian, the *kyrios*, and a set of rooms at the back of the house (HC-01). Sparta, a few days' walk away, let women inherit, own and run land, so that by Aristotle's time they held about 40 per cent of it and managed the estates and the helots while the men were in

barracks, which Aristotle thought was ruining the place (HC-01). Rome began with marriage *cum manu*, the wife passing into her husband's hand, and shifted to marriage *sine manu*, the wife staying legally in her father's line and, once he died, *sui iuris*, her own person, which is how Livia and Agrippina held what they held (HC-01). In Han China the bamboo statutes dug up at Zhangjiashan, compiled around 196 to 186 before Christ and read by Anthony Barbieri-Low and Robin Yates in 2015, show a state that wanted small households for tax purposes and so let women head them and inherit, and preserve a case in which a Widow Wan sued her husband's male relatives for her property and won (HC-01). In Japan the Heian court of 794 to 1185 practised uxorilocal marriage, the husband moving to or visiting the wife's house, as Wakita Haruko showed in 1992, and then the warrior household system of roughly 1100 to 1600 stripped that away until, by the Edo period, little was left (HC-01). Early Vedic India, in Stephanie Jamison's 1996 reading, could not perform its great rituals without the chief wife, and remembered Gargi and Maitreyi as women who argued philosophy with men, before the law books of the Manusmriti filed the memory away (HC-01).

The lever, in every case where the lived column was fuller than the written one, was property. Nowhere is this clearer than in Yossef Rapoport's 2005 study of Mamluk Egypt and Syria. The law gave a daughter half a son's share, and gave the husband the right to divorce. It also gave the bride the *mahr*, a sum paid to her and not to her father, which stayed hers throughout the marriage, and which the husband forfeited in part if he divorced her, while she could buy her own way out by *khul'* (HC-01). The result was a divorce rate in the medieval Islamic cities that Rapoport puts at approaching 30 per cent (HC-01). In England, by contrast, under coverture, which William Blackstone later put into its classic form, a wife's legal person was folded into her husband's, and the only way round it was to be a *feme sole*, an unmarried woman or a trader by town custom, until the Married Women's Property Acts of the late nineteenth century (HC-01). Iceland's *Grágás* barred women from the Thing and shortened their inheritance, and the same women ran the farms, carried the keys of the house at their belt, and could divorce by declaration before witnesses (HC-01).

Then the exceptions. The kingdom of Dahomey, in Edna Bay's 1998 *Wives of the Leopard*, had women's regiments, the Mino, and a reigning

queen mother, the *kpojito*, as an office of state (HC-01). Egypt coded the office of pharaoh as male, so that Hatshepsut, to hold it, had herself carved with the false beard and the kilt: a woman with absolute power in a job that stayed, conceptually, a man's (HC-01). The Haudenosaunee, whom the French called Iroquois, were matrilineal, and the Clan Mothers chose the sachems who sat in council and could remove them (HC-01). None of these dissolved the categories. They drew the line somewhere else.

Why did the line get drawn where it was drawn, and why has it moved? Chapter 2 gave the plough. The research of the last decade has given the herd. Anke Becker, in the *Review of Economic Studies* in 2024, traced more than 1.3 million people in over 100 countries to the subsistence of their ancestors and found that descent from pastoralists predicts restricted movement for women, infibulation, and present-day opposition to abortion, the argument being that a man away with the animals could not watch his wife and so his culture evolved rules that did it for him (SO-05). And Jonathan Schulz, with Bahrami-Rad, Beauchamp and Henrich in *Science* in 2019, argued across 146 nations that the Western Church's bans on cousin marriage, polygyny and arranged marriage dissolved Europe's clans, measured on a Kinship Intensity Index, and so produced the strange, individualistic, nuclear-family West (SO-05). The Church thesis is contested, and its best-known critic is poorly sourced, so take it as a hypothesis rather than a finding.

Persistence has a model of its own. Paola Giuliano and Nathan Nunn, in 2021, divided the years 500 to 1900 into seventy generations of twenty years and measured how much the climate swung from one to the next, and found that where it stayed stable, so did custom, because what the grandparents did still worked (SO-05). Where the plough and the herd and the clan kept paying, the rules they had made kept being obeyed. The weakening, when it came, came from two directions at once. Matthias Doepke and Michèle Tertilt asked in *Econometrica* in 2009 why nineteenth-century men, who held every vote, gave women property and then political rights, and answered that once education paid, men wanted household money spent on children's schooling, and women, on the evidence, spent it that way (SO-05). Claudia Goldin and Lawrence Katz in 2002 used the state-by-state spread of American laws letting unmarried women under 21 obtain the pill to show that early access cut motherhood before 22 by 14 to 18 per cent and raised employment at ages 26 to 30

by about 8 per cent (SO-05). My mother was born in 1946 and I in 1975, which puts the two of us on either side of that hinge.

Note the layer. Every one of these is a cultural explanation. But the plough, the herd and the pill are all answers to a body: to the upper-body strength the plough wanted, to the paternity the herder could not see, to the pregnancy the pill made optional. Culture here is physiology with a long memory.

What the scriptures filed

Between the doctors and the lawyers sits a thinner folder, and it is the one most people reach for first: what the religions and philosophies said the two sexes were *for*. The doctrine, almost everywhere, is complementarity. Two halves, each incomplete, together whole. The layer is culture, though every tradition in the folder insists it is describing the body.

Genesis has God make the woman as a helper for the man, and the English word "helper" carried a thousand years of subordination. The Hebrew is *'ezer kenegdo*. David Freedman pointed out in 1983 that *'ezer* occurs 21 times in the Hebrew Bible and in most of them it is God, coming to rescue Israel, so that the phrase reads less as "assistant" than as "a strength that stands facing him" (HC-02). The rabbis who inherited that text did something the Greek-minded church did not. Daniel Boyarin's 1993 *Carnal Israel* shows a tradition that refused to renounce the body, and the Talmud, needing to assign religious duties, found that two sexes did not cover the people in front of it and used six words: *zachar* and *neqeva*, male and female, *androgynos* for a person with both sets of organs, whom Rabbi Meir called a creature of its own kind, *tumtum* for a person whose sex could not be told, *ayelonit* for a woman who did not develop, *saris* for a man who did not (HC-02). That is a legal system filing physiological variation without pretending it was not there. The Sambia of New Guinea, whom Gilbert Herdt described in 1994, had a word, *kwolu-aatmwol*, for children born looking female who turned male at puberty, which is what a deficiency of the enzyme 5-alpha reductase does, and the word was there before the enzyme was (DH-10). A fact of the physiological layer, given a cultural name. The three layers, filed correctly, by people with no idea they were doing it.

The Christian line ran from Aquinas, for whom grace perfects nature and the sexes complete each other, to John Paul II, whose Theology of the Body was delivered as 129 Wednesday audiences between 1979 and 1984 and remains the most elaborate modern statement of complementarity as a doctrine, and which Patrick Snyder and other critics read as hierarchy with the manners improved (HC-02). Further east the symbol is kinder and the use of it was not. The Taijitu, the yin-yang circle, puts a dot of each inside the other: no one is all one thing (HC-02). The Han philosopher Dong Zhongshu took the same circle and made it a floor plan, *nei* and *wai*, inner and outer, and assigned the wife the inside (HC-02). The Shiva of the Ardhanarishvara image is male down the right side and female down the left, and the Tantric traditions that honoured it used the equality of the halves to break caste and purity rules (HC-02). The Bugis of Sulawesi, in Sharyn Graham Davies's account, have five genders, *makkunrai* and *oroané*, women and men, *calalai* and *calabai*, who live in the other's role, and the *bissu*, who are priests and neither (HC-02).

What can be tested is what the doctrine does when it lands in a questionnaire. Peter Glick and Susan Fiske, in 1996 in the *Journal of Personality and Social Psychology*, built the Ambivalent Sexism Inventory from six studies and 2,250 respondents and found that sexism has two faces that go together, a hostile one, which dislikes women who compete, and a benevolent one, which cherishes women who do not, and that the two travel together (HC-02). Benevolent sexism is complementarity with the theology taken out and the manners left in. A dot of yin in the yang is a beautiful idea. It has also been, for about two thousand years, a reason to keep the yang in the outer room.

The last unsettled fight

Now the top folder, the one still being written to, and the only part of the drawer where the three layers are fighting each other in public. Chapter 5 gave the headline numbers and one fight. Here is the full file.

The claim, stated in the behavioural layer where it lives, is this: on a range of self-reported traits and preferences, the average gap between men and women is larger in richer and more sex-equal countries than in poorer and less equal ones. In David Schmitt's 55-nation personality

study of 2008 the gaps rose to 0.70 of a standard deviation in the most equal nations (DH-10). Erik Mac Giolla and Petri Kajonius found in 2018, across about 130,000 people in 22 countries, that a nation's equality index correlated 0.69 with the width of its personality gaps (HC-07). Shalom Schwartz and Tammy Rubel, in 2005, pooled 127 samples and 77,528 people from 70 countries on basic values and found men higher on power and achievement, women higher on benevolence, at a median gap of 0.15 and a maximum of 0.32, and found the gaps wider where equality was higher (HC-07). Armin Falk and Johannes Hermle, in *Science* in 2018, took the Global Preferences Survey, 80,000 people in 76 countries covering about 90 per cent of the world's population and income, and six economic preferences measured by both questionnaire and money choices (HC-07). Worldwide, women were less willing to take risks, 0.168 of a standard deviation below men, less patient by 0.050, more altruistic by 0.106, more positively reciprocal by 0.055, and more trusting (HC-07). Those are small gaps with wide overlap. But their width correlated 0.67 with the log of a country's income per head and 0.56 with its equality index, which is to say they were near zero in the poorest countries and largest in the richest (HC-07). The paradox, on the preference data, is not that women and men differ much. It is that they differ most where they are freest to be the same.

The cognitive folder reads differently. Luigi Guiso and colleagues, in *Science* in 2008, used the 2003 round of the PISA tests across 40 countries to show that the boys' advantage in mathematics shrinks to nothing in the more equal countries while the girls' advantage in reading grows (HC-07). Gijsbert Stoet and David Geary built on that in 2018 and found that in nearly every country more girls were capable of a university science course than enrolled in one, and the gap between capability and enrolment was widest in the most equal countries, a result they have since extended to 2.5 million adolescents across the five PISA rounds from 2006 to 2018 (HC-07). The first column's reading is that when a girl in Helsinki is free to follow what she likes, she follows it, and what she likes, on average, is not engineering.

Now the other column, and the pages chapter 5 left out. Thomas Breda and Clotilde Napp, in the *Proceedings of the National Academy of Sciences* in 2019, with 300,000 pupils in 64 countries, measured something no cross-country comparison had: the gap inside one pupil between

her reading score and her mathematics score (HC-07). That gap is 0.80 of a standard deviation larger for girls than for boys. A girl who is excellent at mathematics is, more often than a boy of the same mathematics, even stronger at reading. Control for that one within-person difference and the gap in intention to study mathematics falls by 75 per cent, and the gaps in mathematical self-concept and declared interest disappear (HC-07). Brian Nosek and colleagues, in the same journal in 2009, had over 500,000 implicit association tests from 34 countries, found 70 per cent of people pairing science with male faster than with female, and found that the national strength of that pairing predicted the national science gap (HC-07). Breda, with Elyès Jouini and Napp, in 2020, went back to PISA and found that in 71 of 72 countries girls were more likely than boys of the same ability to blame a failure on lack of talent, and that the gap was largest in the wealthy OECD countries (HC-07). The girl in Helsinki, on this reading, is not following what she likes. She is following what she has been told she is like, and the telling is louder in Finland than in Tunisia.

Then the review that should have settled it and did not. Agneta Herlitz and colleagues, in *Perspectives on Psychological Science* in 2024, read 54 papers and reanalysed 27 large studies and found that the paradox is real and that it is particular (HC-07). In richer and more equal countries the gaps grow for personality, verbal ability, episodic memory and negative emotion. They shrink for sexual behaviour, for what people want in a partner, and for mathematics (HC-07). Whatever is widening the one set is not widening the other. The two large theories in the field, the one that says freedom reveals a physiological difference and the one that says prosperity breeds an essentialist culture, each explain the half of the list the other cannot.

So here is the honest sentence, and it is the sentence the Preface promised. The pattern is robust, it is behavioural, and it is cross-sectional. The mechanism is unknown. Not disputed, not contested, not "complex": unknown, in the plain sense that the two best explanations make opposite predictions about the stereotypes, the stereotype data favour one, the preference data favour the other, and nobody has yet run the study that would decide it, which would have to follow the same people in the same country as that country changed. Anyone who tells you the paradox proves what women are really like is handing you a folder. Anyone who tells you it has been debunked is handing you a different folder. The drawer holds both.

What it hands you, instead, is the tool the rest of this book has been assembling. Every page in the drawer can be filed on one of three shelves. The physiological shelf, where menstruation and lactation and the 5-alpha reductase child go, and where the old doctors, for all their rennet, mostly got the observations right. The behavioural shelf, where the paradox lives, as an average with an overlap and a cross-sectional correlation and no mechanism. And the cultural shelf, where the plough and the herd and the Church and the pill go, and the floor plan, and the helper, and the girl's belief about her own talent. The civilisations never had the shelves. They had one drawer and put everything in it and then argued about the drawer. You have the shelves. The next chapter, which is about what to build with them, calls the result Emancipation 2.0, and it begins at the scale this one ends on, the world.

THE STRONGEST CASE AGAINST

The chapter's quiet thesis, that something stable in the body shows through the variety of custom, has two serious opponents, one from each end of the drawer.

The first is the modern one, and it is the one this chapter has to meet head-on. Breda, Jouini and Napp's 2020 result is not a quibble about method. It is a prediction that the first column failed. If widening gaps in rich countries were freedom revealing a disposition, the stereotypes about that disposition ought to weaken as the countries grow freer, because fewer people would need them. They strengthen. In 71 of 72 countries girls blamed their failures on talent more than equally able boys did, and the gap was largest in the richest OECD states (HC-07). The 80 per cent within-pupil reading advantage that Breda and Napp found in 2019 removes three quarters of the intention gap before any disposition is invoked (HC-07). Add the measurement critiques. Ville-Juhani Ilmarinen and Jan-Erik Lönnqvist argued in 2024 that computing the sex difference as a score and then correlating that score with equality overstates the relation by construction, so that part of the paradox is arithmetic (HC-02). The reference-group objection, which HC-07 reports without a primary source, says that a woman in a tight society rates her own assertiveness against other women and a woman in a loose one against everyone, so the

questionnaire is measuring the comparison group, not the trait (HC-07). Andrey Formozov argued that once general development is controlled for, the separate effect of the equality index in the preference data is fragile and may be an artefact of how self-report scales translate (HC-01). And Fors Connolly and colleagues, as chapter 5 reported, found that when a single country grows more equal over time its own gaps do not widen (HC-07). Put together: the gaps may be real, and the paradox a fact about which countries are rich, which is a fact about history, not hormones.

The second opponent is Laqueur, and he survives his own defeat in a weaker form. Grant that Platter drew a female pelvis in 1583. The words for the body were still, always, cultural, and the chapter's own medical folder shows how large the cultural term can be: the same premenstrual syndrome at 12 per cent in France and 98 per cent in Iran (HC-03). If a filter can swing a physiological measure by 86 points, the historian is entitled to ask how much of any cross-cultural sex difference is the thing and how much is the filter, and the chapter cannot answer that for the paradox. Laqueur's critics could not answer it for the skeleton either.

I think the first opponent is winning on the STEM data and losing on the preference data, and I think the second has a point that no amount of data can dismiss. I do not know who is right. I know that the fight is on the behavioural shelf, and that both sides keep reaching for the other two.

IN PRACTICE

The world is the one scale where you cannot run the experiment, so what follows is about reading, not doing. These are conditionals for the person who meets a cross-cultural claim about the sexes, which is to say everyone who reads a newspaper.

If you read that a sex difference is larger in Sweden than in Saudi Arabia, the first useful question is which shelf the claim sits on. If it is a questionnaire about what people say they prefer, it is behavioural, cross-sectional and subject to every objection in the section above, and the sentence that follows it, about what women really want, is not licensed by it. If it is a measured outcome, strength or survival in a famine, it is physiological and the cultural reading has to work harder. The failure mode is the reverse: reading a self-report as if it were a blood test, or a blood test as if it were a preference.

If you are tempted by the freedom-reveals story, because it flatters a side you are on, the test is to ask what it predicts about stereotypes, and then to look. If stereotypes weaken where gaps widen, the story holds. If they strengthen, as Breda found, the story has a hole in it that you are obliged to carry around with you. If you are tempted by the stereotype story for the same reason, the test is the other half of Herlitz's list: a mechanism that explains why girls in rich countries avoid engineering has also to explain why those same countries show larger gaps in episodic memory and negative emotion, which no one is stereotyping anyone about. The failure mode on both sides is the same. It is keeping the half of the list that agrees with you.

If you hire, teach or manage across cultures, and you have in your head a map of which countries produce which kind of woman or man, the map is likely to be a map of averages measured in one year, and the person in front of you is a point, not an average. The overlap on every preference Falk and Hermle measured is wide enough that, in any country, most women and most men sit inside each other's range. The failure mode is the ecological one: treating a country's bell as a person's label. The second failure mode is subtler. If you correct for the first by assuming no differences at all, you are likely to design a workplace that fits the people who already fill it and then be surprised that it does not fill with the others.

If you inherit a norm, a floor plan, a rule about who sits where, and you want to know whether it is a response to anything, the ledger suggests looking for the body it answered. The plough answered a shoulder. The herd answered a husband's absence. If the body the norm answered is no longer in the room, because the plough is a tractor and the pregnancy is optional, the norm is running on memory, which Giuliano and Nunn's model says it tends to do for as long as nothing disturbs it. The failure mode is assuming that a norm with a reason once must still have one, or that a norm without one now never had one. Both lead you to push on the wrong wall.

If you are the one trying to change a norm, the ledger has one more thing to say, and it is not comfortable. In every civilisation where the lived column was fuller than the written one, the lever was property. The *mahr*, the Spartan estate, the *feme sole* trading licence, the ring of keys. Doepke and Tertilt's men gave rights when it paid them to. If a change you want does not change who holds what, it is likely to be a change in the written

column, which is the column history has shown it can leave without effect. The failure mode is believing a declaration.

If you want to use the paradox as evidence for anything, and in a world-scale argument you are going to be tempted, the evidence it supports is narrow. It supports the sentence "cross-sectional sex differences in self-reported traits are larger in richer countries". It supports nothing about why. If someone uses it to tell you what to do about girls in science, or boys in nursing, they are using a correlation with no mechanism as a mechanism, and you are entitled to ask which of the two columns they have chosen and why. The failure mode is the one the Preface warned of: finding the drawer, pulling out your folder, and calling it the drawer.

And if you read the old doctors and the old lawgivers and are tempted to laugh, the lesson of the medical folder is that they saw most of what we see, and filed it wrongly, and that the filing was the part culture did. The failure mode is assuming that the filing we do now is not filing. The girl in Helsinki ticked a box. Somebody chose the box.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. Pre-modern medicine observed sexual dimorphism throughout and the one-sex thesis is rejected by historians of medicine (HC-03). Formal legal subordination rarely meant powerlessness, and the mechanism of lived power was property (HC-01). Patriarchal norms have ecological origins in plough, herd and threat, persist where environments are stable, and weakened with the price of education and the pill (SO-05). The gender-equality paradox is a robust cross-sectional pattern in the behavioural layer, and its mechanism is unknown (HC-07).

Decoration. The girl in Helsinki, who is illustrative. The wandering womb. Hatshepsut's beard. The dot in the Taijitu.

Contested. The mechanism of the paradox, in full (HC-07, HC-01, HC-02). Schulz's Church thesis (SO-05, with a critic poorly sourced). Laqueur's residual point that the words for the body are always cultural (HC-03). The 47.8 per cent PMS figure, given the 12 to 98 per cent spread (HC-03).

NUMBERS TO CARRY

- Women's willingness to take risks sits 0.168 of a standard deviation below men's worldwide, and the width of such gaps correlates 0.56 with a country's gender-equality index: Falk and Hermle, *Science*, 2018, 80,000 people in 76 countries (HC-07).

- The gap inside one pupil between reading and mathematics is 0.80 of a standard deviation larger for girls than boys, and controlling for it removes 75 per cent of the sex gap in intention to study mathematics: Breda and Napp, *PNAS*, 2019, 300,000 pupils, 64 countries (HC-07).
- In 71 of 72 countries, girls were more likely than equally able boys to attribute failure to lack of talent, with the largest gaps in wealthy OECD countries: Breda, Jouini and Napp, *PNAS*, 2020 (HC-07).
- Divorce in the cities of Mamluk Egypt and Syria approached 30 per cent, on the strength of the *mahr*: Rapoport, 2005 (HC-01).
- Early legal access to the pill cut motherhood before age 22 by 14 to 18 per cent and raised employment at ages 26 to 30 by about 8 per cent: Goldin and Katz, *Journal of Political Economy*, 2002 (SO-05).

THE LIBRARY

- HC-01: sex roles across civilisations, from Sparta to the Iroquois, with the formal-versus-lived distinction and the property lever.
- HC-02: complementarity in scripture, philosophy and the Ambivalent Sexism Inventory, plus the Talmud's six categories.
- HC-03: pre-modern medicine from the Hippocratics to Hildegard, the Laqueur dispute, and the modern checks on PMS and lactation.
- HC-07: the gender-equality paradox in full, both columns, with Herlitz's domain list.
- SO-05: why patriarchy arose, persisted and weakened, from plough and herd to the pill (world-scale material only).
- DH-10: Schmitt's 0.70, the Sambia third word, and the ritual record, used lightly.

CHAPTER 16

Emancipation 2.0

It is a quarter to seven on a weekday morning in a flat in a German city, and two people are standing in a kitchen with one cup of coffee between them. One of them is about to leave for the day. The other is about to take a small child to a nursery of a kind that, in the half of the country I was born in, was a rumour. Neither of them has said anything yet. The kettle has. I am not going to tell you which of the two is leaving, and I would like you to notice, before you read on, what your mind just did with that.

Fifty years ago it needed no holding, because there was one answer, and the language had an insult ready for the woman who gave the other one. Today there are two answers, and the two people in the kitchen chose between them, or believe they did. The tax form on the shelf behind them has had a view on the matter since 1958. So has the school timetable, and the nursery's opening hours, and in the western half of the country the payslip, in the year hers overtakes his. None of those views was arrived at by asking the two people what they wanted.

That is the room this book has been walking towards for fifteen chapters, and it is a small room. Everything in the book is in it. The kettle and the coffee are physiology, measurable: two bodies with different hormone ranges, different muscle, different clocks, one of which can lactate and one of which cannot. The two people are statistics: two draws from two bells that mostly overlap, so that whatever you guessed about who is leaving, you were guessing at a population and the person in front of you may be the other tail. The form, the timetable and the insult are culture,

variable: a wall down the middle of this country moved one of those bells by 350 hours a year and left the other where it stood.

My mother's generation asked one question of that kitchen and answered it so loudly that the sentence seemed finished. *Can she?* She can. The sentence had a comma after it, and nothing was ever written after the comma. The second half, "and how would she and he want to, given what they are and what they are not", was the half that could not be asked without sounding like the people who had wanted the old answer. So it was not asked. It went into the drawer, and the kitchen went on being arranged by the tax form.

This chapter is the second half of the sentence. It contains no new fact. Every number in it has been in the book already, with its counter-source, and here they are put to the only use I think they are good for: not a verdict about either person at the counter but seven conditionals, one for each scale the book has climbed, each written as a hypothesis with the evidence against it on the same line, in the grammar the Preface promised and the German border enforced.

Stay in the kitchen. It is the only laboratory the book has left.

What the first half bought

The first half of the question has a receipt, and it is worth reading before the second half is asked, because it is the strongest thing in the book.

A woman can do anything a man can do, in the sense that mattered in 1971: she can earn, hold the job, keep her name, refuse the child, leave the marriage, and, in the eastern half of this country, do all of that while raising the child and out-earning her husband without anyone reaching for a lawyer. The social-security records of half the German workforce say so: about 1,800 hours a year of paid work for an East German mother a generation after the wall, 350 fewer for a West German one, no gap at all for fathers, and 73 per cent of the mothers' gap attributable to the stigma rather than to wages or the price of childcare (Barbara Boelmann, Anna Raute and Uta Schönberg, *AEJ: Applied Economics*, 2025). The pill, on the American evidence, cut motherhood before the age of 22 by 14 to 18 per cent and raised employment at 26 to 30 by about 8 per cent (Goldin and Katz, *Journal of Political Economy*, 2002). Seven years after a first child an

East German mother is back at 70 per cent of her old earnings, a West German mother at 45, under the same laws. The first half was answered, and with a natural experiment so clean that a book which stopped at chapter 13 would be, on its own terms, unanswerable.

Here is the rest of the receipt. The wall moved the mothers. It did not move the fathers, in either direction. The East abolished the breadwinner norm and kept the same woman doing most of the laundry and most of the thinking about the laundry, and in reunified Germany's family panel couples who start symmetrical still tip at the first birth, East and West alike, and do not tip back. The grandmother kept the child alive in 69 per cent of the populations where anyone counted, the father in 32 (Sear and Mace, 2008, 45 populations). And in the richest and most equal countries the gaps in interest, risk and personality did not close. They widened, and after chapter 15 I still cannot tell you why.

That is the receipt, and it has two columns, and my mother's generation read only one of them, which is the column I would have read too. They had a Paragraph 218 to repeal and a husband's signature to abolish, and a question about what she and he would *want*, asked in 1971, sounded exactly like the question their opponents were asking. Chapter 5 gave the reflex its name: if a difference would be unfair were it true, it cannot be true. The moralistic fallacy is not a stupid one. It is the fallacy of people who have been lied to about nature before and have decided, reasonably, not to be lied to again.

So why ask the second half now, when the first is still unfinished in most of the world? Because the kitchen is already designed. That is the thing the drawer hid. The income splitting of 1958 is a design, and it was drawn from an assumption about what a wife is for. The half-day school is a design. The fenced pitch with one gate, the toilet floor plan, the 55-hour week, the sentence 63 per cent longer for the same file, the physical standard written for one skeleton and applied to two: every one of those is an arrangement somebody chose, and most of them were drawn either from a difference nobody had measured or from a sameness nobody had checked. There is no item on the menu called no design. There is design from folklore, design from denial, and design from the evidence, and the first two have had fifty years each.

The books on the subject, mapped for me by the one document in this chapter's library that reads shelves rather than studies, divide the

same way. One shelf explains why the sexes differ and stops at the Pleistocene: Matt Ridley, David Geary, David Buss. The facing shelf says the differences are made and stops at the nursery door: Lise Eliot, Gina Rippon, Angela Saini. Between them Cordelia Fine and Carole Hooven argue about one molecule with Robert Sapolsky holding the ring, Sarah Blaffer Hrdy goes to the crèche and stays there, Richard Reeves goes to the policy desk, and Esther Perel, David Deida and Ivan Illich write about the two in the kitchen as poets, which is honourable and not evidence. The mapping document calls the gap a book that treats the physiology as settled and climbs the scales. I would take the second half of that and leave the first. Nothing in what follows is settled, which is why every principle below carries its own refutation beside it, in the refutation's numbers.

Seven scales, then. At each one, the same three questions the Preface set, and a fourth that the Preface could only promise: what is measurable, what is statistical, what is ours to decide, and, given all three, what arrangement tends to produce what.

The couple, the family, the team

The couple. The design document that seeded this chapter proposes that a stable pair bond needs a heavily lopsided ratio of positive to negative exchanges, roughly five to one, and a daily supply of small bids for attention that the other partner turns towards rather than away. The source is Gottman's observational work of 1994. The document puts the refutation on the next line: Heyman and Slep showed in 2001 that the prediction accuracy Gottman reported, 90 per cent in the document's telling and 93 or 94 in the chapter 6 papers, was fitted to the sample it was tested on, and that applied to new couples the positive predictive value fell to about 29 per cent. Chapter 6 found the five-to-one figure itself traceable, in every document the library holds, to a couples-app blog. So the principle as proposed is a risk factor dressed as a ratio, and I will not carry it.

What survives from the couple's scale is less tidy and more useful. Flooding is measurable: a pulse above about 95 to 100 beats a minute, adrenaline and cortisol up, the frontal cortex off duty, which is physiology and nobody's fault. Who floods is statistical, and the sex that reacts more depends on the instrument. The men flooded faster and stayed flooded

longer on the heart-rate monitor in Gottman and Levenson's 1985 sample. The women showed the larger and longer rise in cortisol and adrenaline in Kiecolt-Glaser's, especially when the husband withdrew. Demand and withdraw correlates with dissatisfaction at r = 0.360 across 14,255 people, and the demander is whoever wants the change, not whoever is the woman.

So the principle, rewritten: given that one body in the argument tends to leave before the person does, a couple that agrees the stop rule before the first flood, and agrees who is asking for change this time rather than who is the woman, tends in the observational data to keep the argument from becoming the pattern. The counter-evidence sits inside the same data. Joel and colleagues' 2020 machine-learning pass found relationship variables explaining 45 per cent of satisfaction and individual ones 21, with the person's sex nowhere near the top, and the best any model could do with change over time was 15 to 20 per cent. If the accent is in the couple at all, it is a quiet one, and the loudest predictors have no sex.

The family. The seed document's second principle is the one this book was built on: a family holds together when provisioning and childcare are spread across people who are not the child's parents, and it cites Hawkes and colleagues' 1998 paper in *PNAS* on grandmothering, with a sample it gives as 51,012. The Grandmother Clause came ashore in chapter 9 with the figures that carry the most weight in the book: across 45 natural-fertility populations the mother's presence improved a child's survival in every study, older siblings' in 83 per cent, the maternal grandmother's in 69, the father's in 32.

Then the father, who in that ledger looks like decoration, and is not. In 89 first-time parents the connectivity between the amygdala and the cortex rose in a straight line with the hours of care a father did, and gay fathers raising a child without a mother showed both the maternal and the paternal network at once (Abraham, Feldman and colleagues, *PNAS*, 2014). The hormone fell with the hours too, 26 per cent at waking and 34 in the evening at the median over four and a half years in Cebu. The father's share in the survival data is small because, in the populations where survival was counted, he was rarely the one holding the child. Where he holds it, the body follows.

Rewritten: given that no human child has ever been raised by two adults alone for long, a family that builds the third adult in, by kin where

the kin are near and by paid care where they are not, tends to do what the pod does, and given that the parental body follows the caregiving hours and not the sex, the parent who takes the hours tends to become the parent the data describe. The counter-evidence is in three places. The seed document's own: market childcare substitutes for kin at a cost nobody has measured well, and chapter 9's question about the cortisol rise in daycare is still open. Krummhörn's: a paternal grandmother in the house raised neonatal mortality 2.48 times in about 32,000 parish records (Voland and Beise, 2002), so "kin" is not one word and the third adult has to be chosen, not assumed. And the plain one: the 32 per cent is not a ceiling anyone has tested in a population where fathers do half the hours, because there is no such population in the ledger yet.

The team. The third principle is the one whose factor fell. Woolley and colleagues reported in *Science* in 2010 that a group has a collective intelligence of its own, weakly tied to its cleverest member and strongly tied to equal turn-taking and to the average score on a test of reading emotion from eyes, on which women score higher, so that teams with more women thought better. The counter is a meta-analysis by Credé and colleagues, dated 2019 in the seed document and 2017 in chapter 10's, and whichever year it was, on 857 groups the single factor extracted 22 to 35 per cent of the variance with an eigenvalue of 1.05 and correlated with task performance at 0.26, which is a factor the way a drizzle is weather.

What survives is a pair of findings about doors and rooms rather than about minds. The room: in 342 four-person audio-only teams, men said almost 70 per cent more words than women (Hardt, Mayer and Rincke, *Management Science*, 2024). The door: when about 7,000 job seekers were offered a job whose pay was competitive, 72.5 per cent of the men and 46.7 per cent of the women applied if applying meant opting in, and 76.4 and 75.3 if the competitive pay applied unless they opted out (Flory, Leibbrandt and List, *Management Science*, 2015). A 26-point gap became a one-point gap when the form changed and the people did not.

Rewritten: given that the measurable sex gap in a team sits at the door and in the airtime and not in the work, a team that changes the form and names the chair tends to close more of its gap than a team that changes its headcount. The counter-evidence cuts both ways. From the biological shelf: the objection chapter 10 carried from Gillen, Snowberg and Yariv (2019), that the entry gap may be risk and confidence measured with error,

so that the preference the form seems to override may not have been a preference at all. From the other shelf: Alice Eagly and Wendy Wood's 2023 finding, in the document that argues against this book, that 71 per cent of meta-analytic sex differences mirror the effects of experimentally assigned power. If airtime tracks power, then naming the chair fixes power, not sex, and the principle works for a reason that has nothing to do with the accent. The two people in the kitchen do not need to know why the form works to use it.

The organisation, the city, the state, the world

The organisation. Here the seed document's principle arrives without a primary source, the first of two that do. The hypothesis is that an organisation which ties advancement to unbroken hours selects, without meaning to, for the sex with the looser claims on its evenings, and bleeds the other. The evidence offered is "research in the financial services sector and broader economic models", a gesture rather than a citation. The long week is a measured killer and it kills mostly men: 55 hours or more a week carried a relative risk of stroke of 1.33 across 528,908 people (Kivimäki and colleagues, *The Lancet*, 2015), and in 2016 long hours were attributed 745,194 deaths worldwide, 72 per cent of them male (Pega and colleagues, 2021). Leadership by sex, meanwhile, differs almost not at all: task orientation at $d = 0.00$ and interpersonal at 0.04 across 162 studies (Eagly and Johnson, 1990). The labelled accommodation is used by nobody and paid for by the people it names: Japan's menstrual leave by 0.9 per cent of the women entitled to it, while 80.7 per cent of 32,748 Dutch women worked through pain at a third less output (Schoep and colleagues, *BMJ Open*, 2019), and a mandated benefit for one group tends, in the labour economics since Gruber in 1994, to come out of that group's wages.

Rewritten: given that the unbroken week damages mostly men's bodies and the labelled leave prices mostly women's hours, an organisation that pays for the outcome rather than the hour, and builds the accommodation nobody has to be named to use, tends to keep more of both sexes than one that does either half. Fontevraud kept monks under an abbess for 689 years because the one thing it put in writing was who

answered to whom. The counter-evidence is partly the seed document's own, that some trades are continuity and the hour is the product, and partly the book's. The kibbutzim set out to abolish the division of labour and within two generations had one again, for reasons the people who ran the numbers still dispute. And where the task is the body, no reward structure moves it: in the Marine Corps task force of 2015, 93 of 134 tasks showed a significant sex difference and 88 of those favoured the all-male teams. Decoupling reward from the hour is a design for the desk, not the stretcher.

The city. The second principle without a source. Build the home, the work and the childcare within reach of each other, the document says, and the unpaid-labour divide loses its friction. Its one number is that women do "20 to 1000 per cent" more unpaid work than men, credited to global time-use surveys by way of an IMF working paper. Chapter 12 found the city carried by one document and said so. It also found a handful of measurements that hold regardless. In Montréal, 28 per cent of all trips are care trips, 32 per cent of women's and 25 per cent of men's (Ravensbergen and colleagues, *Transport Policy*, 2023). A woman takes 178.9 seconds at the toilet to a man's 118.4, so equal floor space produces a queue by arithmetic and a fixture ratio of 2:1 to 3.25:1 removes it (Bovens and Marcoci, 2020). And Vienna, since Eva Kail's exhibition of 1991, has been the one city that counted the loop before it built for it.

Rewritten: given that the care loop in every city measured is walked mostly by women today, a city that counts the loop and builds for the loop, not for the women, tends to shorten the day of whoever walks it tomorrow. The counter-evidence is the plough. Chapter 15 carried Alesina, Giuliano and Nunn's finding that the tool a society farmed with a thousand years ago still predicts who leaves the house today, and a city that makes one arrangement comfortable tends to keep that arrangement, so the loop you build for may be the loop you fix in place. The seed document adds its own: proximity priced by a market becomes a luxury, and the carers commute in from the edge to serve it. And chapter 12's hardest question stands: whether the city is about sex at all, or about care, which any adult can do and which the Endocrine Accent has almost nothing to say about.

The state. The sixth principle is the one the Preface promised: across 175 countries and more than 130,000 data points in the WomanStats data-

base, the best single predictor of a state's peacefulness is the physical security of its women, ahead of its wealth and its democracy (Hudson, Caprioli, Bowen and McDermott, *Sex and World Peace*, 2012). The same document adds that legal polygyny breeds a surplus of unmarried young men and instability, and here the book has a hole it will not paper over. The research document commissioned on sex ratios and polygyny was never produced, chapter 14 marked the gap, and I leave the clause as an open question rather than fill it from memory.

What chapter 14 did establish is why the principle must be read as a fact about the state and not as a promise about the women. Across 193 European reigns from 1480 to 1913 a polity under a queen was 39 percentage points more likely to be at war in a given year than one under a king (Dube and Harish, *Journal of Political Economy*, 2020). A ten-point rise in women's share of an Indian state legislature cut neonatal mortality by 2.1 percentage points, a third of the baseline (Bhalotra and Clots-Figueras, 2014). And under identical Israeli training, stress fractures ran at 23.9 per cent in women against 11.2 in men, and 12.0 against 1.3 in light infantry (Finestone and colleagues, 2005 and 2014).

Rewritten: given that a state's peace tracks the safety of its women and not their supposed gentleness, a state that designs for the safety tends to get the peace, and a state that designs for the gentleness tends to get the queens. The counter-evidence is in the direction of the arrow: the critics chapter 14 carried hold that a state which violates its women may be unstable for a third reason that produces both, and no multivariate model has yet closed that door. On the standards, the counter cuts both ways: a sex-blind standard breaks women's bones at double and nine times the rate, and a sex-lowered standard breaks the trust of the men beside them, so the design is a standard set to the task and not to the sex, with the training built to the bone.

The world. The last principle: lasting settlements and treaty adherence depend on women holding substantive negotiating power rather than an observer's chair, credited to Caprioli in the *American Political Science Review* of 2000, with the Accra talks on Liberia in 2003 as the example. The document's own counter is fair: a woman at the table does not resolve the scarcity that caused the war, and a woman who reaches the top of a masculinised ministry may have got there by adopting its habits.

What the world scale actually hands the chapter is the finding I like

least and trust most. Across 80,000 people in 76 countries, women's willingness to take risks sits 0.168 of a standard deviation below men's, and the width of that gap and five others correlates at 0.56 with the country's gender-equality index: the more equal the country, the wider the gap (Falk and Hermle, *Science*, 2018). The best counter-study, in the document that argues against this book, finds the stereotype that mathematics is not for girls *stronger* in richer countries across 300,000 pupils in 64 countries, and the paradox gone once it is controlled (Breda, Jouini, Napp and Thebault, *PNAS*, 2020). The within-country test found no divergence in values as equality rose across 235,339 people in 32 countries (Fors Connolly, Goossen and Hjerm, 2020). Three good studies, two shelves, no verdict.

Rewritten: given that the measured gaps between the sexes have widened, not closed, in the countries that opened the most doors, and that nobody can yet say whether that is freedom or a subtler stereotype, a world that designs for equality of the door tends to get more difference at the table, not less, and should budget for the difference rather than read it as the door having failed. The counter-evidence is the whole of the facing shelf, and the honest sentence is the one chapter 15 ended on. I don't know. Neither does anyone else yet.

What she would recognise

My mother was born in 1946 and died in 2021, and in between she watched the first half of the question get answered. I think she would recognise more of the second half than she expected to.

She would recognise that every principle above begins where hers did. The state's peace begins with a woman's safety. The organisation's design begins with a woman's hour not being priced. The team's begins with a form that does not quietly ask her to volunteer for the competition the man was enrolled in. The city's begins with counting the trips she makes. The first half was not wrong. It was a prerequisite. You cannot ask two people what they want from a kitchen until both of them are allowed to leave it.

She would not recognise three things, and they are the three the drawer held. The first is that the father is in the design, not as a second mother

but as a body that changes with the hours it does. The second is the grandmother, who had been in the design for two hundred thousand years and was removed from it, in the West, by a tax form and a half-day school that assumed a mother at home and no one else. The third is that the second half of the question has no arrangement as its answer. It has a grammar. Given what we are, under these conditions, this tends to produce that. It does not say what she or he should want. It says what tends to happen when they want it, and it leaves the wanting to them.

That is Emancipation 2.0, and it is not a second wave. It is the rest of the sentence her generation started, written in the only grammar that survived the border.

So the tool. At every scale the book has climbed, the same four questions, in order, and never the fourth before the first three. What is measurable here, and in which body? What is statistical here, and how far do the bells overlap? What is culture here, and who drew it, and when? And then, only then: given all three, what arrangement tends to produce what, for whom, and what is the evidence against it?

Ask them of the kitchen. Then ask them of the two people in it, which is a different and a shorter conversation, because the two people are not the distribution. They are one draw each, and the only thing the whole book can tell them for certain is that nobody else can tell them what to want.

THE STRONGEST CASE AGAINST

Rule three applies to the last chapter as to the first, and here the case is against designing anything from these differences at all.

It goes like this, in the voice of the document that was asked to steelman the other side and argued it instead. After correcting for the 10 per cent size difference, sex explains about 1 per cent of the variance in brain structure (Eliot and colleagues, 2021). Adult hormone levels are not a cause waiting to be designed around but a readout of the life being lived: care lowers testosterone, authority and winning raise it, in both sexes (van Anders and colleagues, *Psychoneuroendocrinology*, 2011). Of the sex differences the meta-analyses find, 71 per cent mirror the effects produced when experimenters hand one group power over another (Eagly and Wood,

2023). The male advantage in negotiation, across 10,888 people, disappears or reverses when women bargain for someone else, when the range is transparent, or with experience (Mazei and colleagues, *Psychological Bulletin*, 2015). Caregivers lay gender into infants between birth and 15 months through how they are held and spoken to (Fausto-Sterling, 2012). And the wall: 73 per cent of the German mothers' gap is the stigma, and none of it is the wage (Boelmann, Raute and Schönberg, 2025).

From those numbers the case draws a conclusion sharper than any the Preface faced. It is not merely that the accent is small. It is that every design drawn from a difference manufactures the difference it was drawn from. The income splitting of 1958 was drawn from an assumption about wives and produced, two generations later, wives who fit it. The plough was drawn from upper-body strength and produced, a thousand years on, a norm about who leaves the house. If culture becomes biology by fifteen months, then a fixture ratio, a care loop built into a transport network, a crèche sited for the person who does the care, are stereotypes poured in concrete, and the person who walks the loop tomorrow finds it already waiting because this book told the council to build it. On this reading the only honest design is the one that assumes nothing about either sex, measures the individual, and lets the bells fall where they fall. Anything else is the naturalistic fallacy with a planning permit.

I think that case is right about more of this chapter than I would like. Count the seven principles and ask which layer each one stands in. The couple's stop rule, the team's form, the organisation's unlabelled accommodation, the city's loop, the state's safety: those are culture-layer designs, and every one of them would be worth building if the Endocrine Accent were silent, which is why I wrote them for the hour and the loop and the form, and not for the women. Only three things in the chapter lean on the body: the gestation and lactation no agreement moves, the bone under the pack, and the stretcher. The case against is also right that a design which names a sex re-makes it, as the labelled leave used by 0.9 per cent showed.

Where I still part from it is the kitchen in the East. The constructivist case predicts that where the breadwinner norm was abolished for forty years and the stigma ran the other way, the laundry should have moved. It did not. The fathers did not move in either Germany, under either insult. Something in that stillness is either a second wall nobody has found or an

accent nobody has measured, and a design that assumes nothing about it is still a design, drawn from a sameness nobody has checked. The case against has the stronger numbers. It does not yet have that kitchen.

IN PRACTICE

This is the last chapter, so the scale is yours, whichever it is. What follows is one conditional for each, with the way it fails and the number against it, and a last one for the reader of this book.

If you are one of the two people in the kitchen, and you are about to decide who leaves at seven, then write the decision in the book's grammar before the first birth and not after it: given our two jobs, our two bodies and the hours the nursery keeps, this arrangement, until this date, and we look at it again then. The failure mode is the one chapter 7 found in the panel: the decision made in month seven of the pregnancy tends to be overruled in month three of the baby by lactation, which is the one bottleneck no agreement moves. The counter-evidence is the eastern half of the same panel, where the same bodies tipped differently, which tells you everything after the milk is negotiable.

If you catch yourself using an average on the person across the table, then stop and do the arithmetic once: at $d = 0.21$ the bells share about 85 per cent of their area and the random woman outscores the random man 42 times in a hundred. The failure mode runs the other way too. The same overlap used to deny the average when you are planning for a thousand people is the other jaw of the trap, and the multivariate distance of 2.71 on 10,261 people, contested as it is, is the counter-evidence that the thousand are not interchangeable even when any two of them might be.

If you are building a family, then build the third adult in before the first child arrives, by kin where the kin are near and by paid care where they are not, and if it is kin, choose which. The failure mode is Krummhörn: a paternal grandmother in the house raised neonatal mortality 2.48 times in one parish, and the maternal grandmother lowered it, so "family" is not a design and a named person is. The counter-evidence is that the cost of paid care to a child, in cortisol and in what else, is chapter 9's open question and still open.

If you run a team and you want the gap closed, then change the form and name the chair before you change the headcount. The failure mode

is that the form fixes the door and not the room, and in the room the men tend to say 70 per cent more words. The counter-evidence is that the gap at the door may be risk and confidence measured badly rather than a preference at all, in which case the form has corrected a measurement and not a person.

If you run an organisation, then pay for the outcome and not the hour, and build the accommodation nobody has to be named to use, the window, the moved shift, the day role on request, for everyone, including the man with the migraine. The failure mode is the senior man's own week: a policy tends to be the hours the most senior man is seen to keep, and the 55-hour week is a measured killer of men before it is anything else. The counter-evidence is the stretcher, where 88 of 93 task differences favoured the all-male team and no reward structure moves a load.

If you sit on a council, then count the care loop before you change a route, and build for the loop rather than for the women who walk it today. The failure mode is the plough: a city that makes one arrangement comfortable tends to keep it, and the loop you build for may be the one you fix. The counter-evidence is one document, one survey in Montréal, and a press release from Vienna, which is the state of the library at this scale.

If you vote, draft or judge in a state, then design for the safety of its women and not for their gentleness, and set every physical standard to the task and the training to the bone. The failure mode is the peace movement that puts a woman in charge of the war ministry and is surprised, with 39 percentage points of queens as the record. The counter-evidence is the arrow: a state that violates its women may be unstable for a third reason that produces both, and the polygyny half of the claim rests on a document this library never got.

If you are designing for the world, which some readers are, then budget for difference under equality rather than reading it as the door having failed. Across 76 countries the gaps widened as the doors opened. The failure mode is the verdict: reading the paradox as proof that the accent is loud, when the best counter-study finds the stereotype louder in exactly those countries. The counter-evidence is Breda's 300,000 pupils, and Fors Connolly's 235,339 people who did not diverge within their own countries as equality rose.

And if you are reading this book, then apply its method to it. Every

figure above is still open, checked against a machine-written library and not yet against the paper. The failure mode is mine: the narrator has a side, a sex and a generation, and a book about men and women that cannot lose an argument to both tends to be read as a prosecution. The counter-evidence is the facing shelf, which has better numbers than my side on the brain, the hormones and the negotiation, and which I have tried to give its own voice in every chapter. Where I failed, the notes at the back are there to catch it.

And if none of this fits the two of you in the kitchen, then the overlap is doing its job, and the question my mother's generation asked is yours to finish. Not whether she can. That half is done. *How would she and he want to.* It is a quarter to seven. The kettle has already answered. The two of you have not.

LOAD-BEARING, DECORATION, CONTESTED

Load-bearing. The three layers and the conditional grammar, from the Preface and chapter 5. The two-column receipt of the first half: the wall moved mothers 350 hours and fathers not at all (Boelmann, Raute and Schönberg 2025, SY-06 via chapter 13); the reversion at the first birth (FReDA via the Preface and chapter 7); the parental body following the hours (Abraham and Feldman 2014, Gettler 2011, chapter 8); the grandmother at 69 per cent and the father at 32 (Sear and Mace 2008, chapter 9). Heyman and Slep 2001 as the reason the couple's principle is a risk factor, not a ratio (SY-05, chapter 6). Flory 2015's form (chapter 10). Kivimäki 2015 and Pega 2021 on the hour (chapter 11). Bovens and Marcoci 2020 and Ravensbergen 2023 on the loop (chapter 12). Dube and Harish 2020 as the reading rule for Hudson 2012 (chapter 14, SY-05). Falk and Hermle 2018 against Breda 2020 and Fors Connolly 2020 as the unresolved pair (chapter 15, SY-01). The constructivist case as the case against design itself: Eagly and Wood 2023, van Anders 2011, Mazei 2015 (SY-01).

Decoration. The kitchen at a quarter to seven. The comma after the first half of the sentence. The shelf map (SY-03's shape, no citations carried). The 689 years at Fontevraud. The drizzle.

Contested. Whether Gottman's accuracy is 90 (SY-05), 93 or 94 (CO documents) in sample, and whether the 29 per cent is positive predictive value or overall accuracy. Credé's meta-analysis, 2017 (OR-01) or 2019 (SY-05). Hawkes 1998's sample of 51,012 (SY-05 only). Hudson 2012's 175 countries and 130,000 data points (SY-05 only, secondary sources). Caprioli 2000 and the Accra 2003 example (SY-05 only, citation mismatch). The IMF's 20 to 1000 per cent (SY-05 only). The arrow in the Hudson correlation. Whether the fathers' stillness in the East is a second wall or the accent (SY-06's open question, chapters 7 and 13, against SY-01's whole case). The

gender-equality paradox, in full. Everything that belongs in the missing SO-02 report. Stoet and Geary's sample, 472,242 in 67 countries (HC-07, SY-03) against over 300,000 in 73 (SY-05). Lassek and Gaulin's muscle figure, about 60 per cent (Preface, via its sources) against 65 (SY-05).

NUMBERS TO CARRY

1. About 29 per cent: the positive predictive value of Gottman's divorce prediction on new couples, against 90 per cent in the original sample (Heyman and Slep 2001, per SY-05. The chapter 6 documents give the in-sample figure as 93 or 94).
2. 69 against 32: the share of 45 natural-fertility populations in which a maternal grandmother's presence, or a father's, improved a child's survival (Sear and Mace 2008, *Evolution and Human Behavior*, per FA-03 and DH-08 via chapter 9).
3. 76.4 and 75.3 per cent: men and women applying for a competitively paid job when competitive pay applied unless declined, against 72.5 and 46.7 when it had to be chosen, about 7,000 job seekers (Flory, Leibbrandt and List 2015, *Management Science*, per OR-03 via chapter 10).
4. 745,194 deaths in 2016 attributed to working 55 hours or more a week, 72 per cent of them men (Pega et al. 2021, *Environment International*, per OR-05 via chapter 11).
5. 0.56: the correlation between a country's gender-equality index and the width of its sex gaps in six economic preferences, 80,000 people in 76 countries (Falk and Hermle 2018, *Science*, per HC-07 via chapter 15), against the paradox vanishing once stereotype strength is controlled across 300,000 pupils in 64 countries (Breda, Jouini, Napp and Thebault 2020, *PNAS*, per SY-01 and HC-07).

THE LIBRARY

- SY-05 Toward design principles for living together: the seven hypotheses at seven scales, with their stated counter-evidence; two of the seven name no primary source.
- SY-01 Steelman the opposite thesis: the constructivist case, carried here as the case against designing from difference at all.
- SY-03 Map of existing books and the gap: the shelf map, used for shape only, no citations carried; its "settled science" conclusion not adopted.
- Chapters 00 to 15, by their own libraries, for every figure not supplied by the three documents above: HC-05, SY-06 (the wall); CO-01, CO-04 (the couple); FA-01, FA-03, DH-08, BI-02, BI-06 (the family); OR-01, OR-03 (the team); OR-02, OR-04, OR-05, OR-06, OR-07 (the organisation); SO-03 (the city); SO-01, SO-04 (the state); HC-07, SO-05 (the world); PS-01, BI-11 (the bells); SY-02, SY-04 (the trap).
- SO-02 Sex ratios, polygyny and social stability: missing, see chapter 14.