THE PATCHWORK

Money, Speed, and the Institutions That Couldn't Keep Up

Z.ai

THE PATCHWORK

Money, Speed, and the Institutions That Couldn't Keep Up

Z.ai

THE PATCHWORK

Money, Speed, and the Institutions That Couldn't Keep Up

First Edition

2026

For educational and illustrative purposes. Some names, dates, and details have been rendered as narrative nonfiction based on public reporting and regulatory filings.

CONTENTS

Chapter 1

The Run That Took Six Hours

On the morning of Thursday, March 9, 2023, a man in a small office on Sand Hill Road in Menlo Park sent an email that, in the ordinary course of business, would have been read by perhaps two dozen people. The man worked at Founders Fund, the venture capital firm started by Peter Thiel, and the email was addressed to the founders of the companies in the fund's portfolio. The subject line was simple. The instruction was simple. The email said, in substance: get your money out of Silicon Valley Bank.

The email, like most emails, was forwarded. By ten o'clock in the morning Pacific time, screenshots of it had appeared in at least a dozen private Slack channels populated by founders whose companies were not in Founders Fund's portfolio. By eleven o'clock, the screenshots had been forwarded again, into channels whose membership ran into the hundreds. By noon, founders across San Francisco, New York, Austin, and Boston were on the phone with their CFOs, asking the same question the Founders Fund email had implicitly raised: was Silicon Valley Bank safe?

Silicon Valley Bank, at the time, was the sixteenth-largest bank in the United States, with one hundred and seventy-three billion dollars in deposits. It was, by a substantial margin, the most important financial institution in the American technology industry. It held the operating accounts of roughly half the venture-backed startups in the country. It held the personal banking relationships of many of their founders. It held the funds that venture capital firms had raised from limited partners and were deploying into companies. It was, in the idiom of the industry it served, the bank of the valley.

By the end of the day, forty-two billion dollars had walked out the door.

The withdrawals did not happen in the way that bank runs are supposed to happen. There were no lines around the block. There were no tellers counting out hundred-dollar bills into shaking hands. There were no depositors camping overnight in folding chairs, the way depositors at IndyMac had done in Pasadena in 2008. The lobby of the bank's branch on Sand Hill Road was, by all accounts, quiet. The lobby of the bank's branch in San Francisco's South of Market neighborhood was, by all accounts, quiet. The withdrawals happened on phones. They happened in Slack. They happened through a wire-transfer button in a mobile banking app, in a transaction that took perhaps thirty seconds to authorize and that the bank was, structurally, unable to refuse.

The FDIC had been designed in 1933, in the depths of the Great Depression, to handle bank runs. The agency's founding insight was that bank runs happened because depositors lost confidence faster than banks could restore it, and that the government could restore confidence more credibly than any individual bank could. The mechanism was insurance: the FDIC would guarantee small deposits, so that even if a bank failed, ordinary depositors would not lose their money. The mechanism worked, for ninety years, on the assumption that depositors, having been insured, would not panic. It also worked on the assumption that, if they did panic, the panic would be slow enough that the FDIC could arrive, close the bank, and arrange an orderly transfer of its assets and liabilities to a healthier institution over a

weekend.

What came next was a bank run, in the technical sense. In the temporal sense, it was something else entirely. The run on Silicon Valley Bank took roughly six hours.

The FDIC, which had been monitoring SVB since Wednesday afternoon, when the bank had announced a capital raise that spooked the market, was technically aware that something was wrong. By Thursday evening, the agency's staff was on the phone with the Federal Reserve, with the Treasury, and with the California Department of Financial Protection and Innovation, which chartered SVB at the state level. By Friday morning, the California DFPI had closed the bank and appointed the FDIC as receiver. The FDIC's staff arrived at SVB's Santa Clara headquarters, presumably prepared to assess the situation, arrange a transfer, and manage the resolution over the weekend in the orderly manner that the agency had managed hundreds of resolutions before.

They arrived to find a building still staffed, still lit, still operational in every visible respect. The tellers were at their desks. The lobby was empty. The bank had been functionally dead since the previous afternoon.

The FDIC did not, in any operational sense, save Silicon Valley Bank. It did not close it, either, except in the formal sense required to take possession. The agency's staff arrived at the bank's Santa Clara headquarters on Friday morning the way a coroner arrives at a house — to certify a death that had occurred the previous afternoon, and to begin the paperwork.

How this happened is the subject of the present chapter. Why it happened — why a bank that took forty years to build could be emptied in six hours, and why the institution designed to prevent such things arrived too late to matter — is the subject of this book. SVB was managed badly. Its regulator was slow. These things are true, and they will appear in what follows. But the bank was not killed by bad management, and the regulator's slowness was not, in the end, the cause of the failure. The cause was a mismatch — between the velocity at which a depositor base, alerted by a screenshot in a Slack channel, could decide to move forty-two billion dollars, and the velocity at which an institution designed in 1933 could respond.

SVB happens to be the case in which this mismatch is easiest to see — the one in which the architecture of the old system and the velocity of the new were both visible at once, in daylight, on the front page of every newspaper. The cases that follow are harder to see, because they happened in places less covered, or because the architecture was invisible, or because the velocity was deniable. They are, however, the same case.

* * *

Silicon Valley Bank was founded in 1983 by two men named Bill Biggerstaff and Roger Smith, who had been playing poker at a San Jose country club when the conversation turned, as conversations among bankers in the early 1980s sometimes did, to the question of why no bank in Northern California

seemed interested in the business of lending to technology companies. The conventional banks of the era — Bank of America, Wells Fargo, the regional banks that served the agricultural and industrial businesses of the Valley before the Valley was called Silicon — regarded technology companies as too small, too speculative, and too unfamiliar to be creditworthy. A hardware startup with three engineers, a prototype, and no revenue was, in the worldview of a loan officer trained in the 1970s, a worse credit risk than a strawberry farmer with a deed to his land. The strawberry farmer, at least, had collateral.

Biggerstaff and Smith, and a third founder named Robert Medearis, raised five million dollars in seed capital and opened the bank's first office in Santa Clara, in a building that had previously been a Chinese restaurant. The bank's business model was straightforward, and remained straightforward for forty years: take deposits from the venture capital firms and the established technology companies that had cash, and lend that cash to the early-stage technology companies that did not. The loans were often secured not by collateral — early-stage technology companies had no collateral — but by the venture firm's commitment to fund the company's next round. If Sequoia Capital had just led a ten-million-dollar round into a startup, SVB would lend that startup two million dollars against the cash that Sequoia had just put in, charging interest and taking warrants that gave the bank a small piece of the upside if the company succeeded.

The model worked, for forty years, because the venture capital business was small enough that SVB's bankers knew, personally, the partners at every major firm. They knew which companies Sequoia had backed. They knew which partners at Andreessen Horowitz were leading which deals. They knew, in the informal way that knowledge moves in a small industry, which companies were going to raise their next round and which were not. The bank's underwriting was, in effect, the bank's social network. SVB was the bank of the valley because the valley was small enough to bank personally.

For most of its history, SVB was a modest institution. In 1995, it had one billion dollars in assets. In 2005, it had five billion. In 2011, when Greg Becker, a career SVB banker who had joined the firm in 1993, was promoted to chief executive, the bank had eleven billion dollars in assets and was a regional curiosity. Becker was thirty-four years old. He had a degree from Indiana University and an MBA from the University of Wisconsin. He had the open, slightly earnest manner of a midwestern banker transplanted to California, and he believed, with some reason, that the technology industry was about to enter a period of extraordinary growth, and that SVB was uniquely positioned to grow with it.

The culture of the bank, in this period, was the culture of the industry it served — informal, optimistic, and slightly contemptuous of the formalities of traditional banking. SVB's bankers did not wear ties. They attended the same conferences as their clients. They sponsored the parties at South by Southwest and hosted wine tastings at the bank's annual gathering in Napa, where venture capitalists and founders would mingle with SVB's relationship managers over Cabernet and small plates, and where the conversation, by all accounts, was less about banking than about the things the bank's clients were building. The bank's marketing emphasized its fluency in the technology industry's particular dialect. SVB did not just lend to startups; it "understood" startups. It did not just hold deposits; it "partnered" with founders. The bank's annual report for 2021, the year of its greatest growth, featured photographs of its clients' products — a robotic arm, an electric vehicle chassis, a stack of code on a

screen — and described SVB as "the financial partner to the innovation economy."

This was not, in the main, marketing fluff. SVB's bankers genuinely did understand the technology industry in a way that the loan officers at JPMorgan Chase and Bank of America, trained to underwrite against collateral and cash flow, did not. A banker at SVB could look at a Series A term sheet and understand, in a way that a traditional banker could not, whether the company was likely to raise a Series B. The bank's credit decisions were informed by an intimacy with the venture capital world that no other bank possessed. The intimacy was the bank's competitive advantage. It was also, in a way that would become clear only in retrospect, the source of the bank's vulnerability — because the same intimacy that let SVB lend prudently to startups also meant that SVB's depositor base was, in a crisis, a single community that would behave like a single community, taking its cues from the same small number of venture firms, all at the same time.

Becker, in any event, was right about the growth. Between 2011 and 2021, the American venture capital industry grew from roughly thirty billion dollars in annual investment to roughly three hundred and thirty billion dollars in annual investment — an elevenfold increase, driven by a combination of low interest rates, the rise of smartphones, the rise of cloud computing, the rise of software-as-a-service, and the arrival of vast pools of capital from sovereign wealth funds, mutual funds, and SoftBank's Vision Fund, all of which had decided, around 2014, that the best place to deploy capital was in private technology companies. SVB grew with the industry it served. By the end of 2019, the bank had seventy-one billion dollars in assets. By the first quarter of 2022, the bank had two hundred and twenty billion dollars in assets, one hundred and ninety-eight billion dollars in deposits, and a market capitalization that briefly exceeded forty billion dollars. Becker, who had been running a regional bank a decade earlier, was now running one of the twenty largest banks in the United States.

The growth was real, and it was, in a sense, the source of the problem. SVB's deposits had grown so fast that the bank could not, structurally, lend them all out. A bank makes money by taking deposits and lending them at a higher interest rate than it pays on the deposits. The spread between the two rates is the bank's net interest margin, and it is the fundamental source of banking profitability. But SVB's depositor base — venture-backed startups that needed their cash for payroll and operating expenses — was not a depositor base that could absorb hundreds of billions of dollars in loans. SVB could lend two million dollars to a startup that had just raised a Series A. It could not lend two billion dollars. The supply of creditworthy early-stage technology companies, in any given year, was finite, and SVB had, by 2021, more deposits than it could prudently lend.

The bank's solution to this problem was the solution that banks have always used when they have more deposits than they can lend: it bought securities. Specifically, it bought United States Treasury bonds and mortgage-backed securities issued by government-sponsored enterprises like Fannie Mae and Freddie Mac. These securities were, by any reasonable definition, among the safest financial assets in the world. They were backed by the full faith and credit of the United States government, or by the mortgage payments of millions of American homeowners. They would not default. They were, in the credit-risk sense, bulletproof.

They were not, however, immune to interest rate risk, and this is the point at which the story of SVB becomes a story about the specific architecture of the bank's balance sheet, and about a decision

that, in retrospect, looks like the decision that killed the bank.

The securities SVB bought were, predominantly, long-dated. They had maturities of five, ten, or thirty years. They paid fixed interest rates — rates that, in 2020 and 2021, when SVB was buying most of these securities, were extraordinarily low. The ten-year Treasury bond yielded around 1.5 percent in 2021. The thirty-year fixed-rate mortgage-backed security yielded around 2 percent. SVB was buying these securities with deposits that, in many cases, were paying zero interest — venture-backed startups do not generally earn interest on their operating accounts — so the spread between what SVB was earning on its securities and what it was paying on its deposits was attractive. The bank was making money, apparently safely, on the difference between 1.5 percent and zero percent.

The catch, which is a catch that every banker learns about in the first week of training, was that the value of a fixed-rate bond moves inversely to interest rates. When rates go up, the value of existing bonds goes down, because new bonds are being issued at higher rates and nobody wants to buy an old bond paying 1.5 percent when they can buy a new bond paying 4 percent. This is not a problem if the bank intends to hold the bond until maturity — at maturity, the bond pays back its face value, regardless of what interest rates did in the interim. The accounting rules, in recognition of this, allow banks to classify certain securities as "held to maturity," in which case the bank does not have to recognize the decline in market value on its balance sheet. The decline is real, but it is unrealized — it exists only on paper, and it will go away if the bank waits long enough.

This is what SVB did. By the end of 2022, the bank held roughly ninety-one billion dollars in securities classified as held to maturity. The face value of those securities was ninety-one billion dollars. The market value was approximately seventy-six billion dollars. The bank had, on its held-to-maturity portfolio, an unrealized loss of roughly fifteen billion dollars. The loss was on paper. It would go away if the bank could wait long enough for the securities to mature.

The problem was that SVB might not be able to wait. The bank's deposit base — overwhelmingly venture-backed startups — was the kind of deposit base that could, in a crisis, disappear quickly. Startups do not keep deposits at a bank because they like the bank. They keep deposits at a bank because they need somewhere to park the cash they have raised from venture capitalists, and they need that cash to be available to make payroll next month. If the bank looks shaky, the startup will move the cash. The deposits are, in the technical sense, "flighty."

Most banks have deposit bases that are not flighty. A traditional retail bank has millions of individual depositors, each with a few thousand dollars in a checking account. Those depositors do not move their money quickly, because moving money is inconvenient and because the amounts are small enough that even a bank failure would not be catastrophic for any individual depositor. SVB had a deposit base of roughly thirty-seven thousand customers, of whom a relatively small number held a very large fraction of the total deposits. By the end of 2022, approximately ninety-four percent of SVB's deposits were above the FDIC's insurance limit of two hundred and fifty thousand dollars. Ninety-four percent of the bank's deposits were, in the event of a failure, uninsured.

This was the structural vulnerability. It was not a fraud. It was not a secret. SVB disclosed its held-to-maturity portfolio, its unrealized losses, and the percentage of its deposits that were uninsured in its public filings with the Securities and Exchange Commission. Anyone who read SVB's 10-K filing

for the year ending December 31, 2022 — and there were analysts who did — could see that the bank had a fifteen-billion-dollar hole in its balance sheet that would become real if the bank were forced to sell its held-to-maturity securities before they matured. The hole was public. The deposit base was public. The mismatch between the two was public.

What was not public, and what would not become clear until March 2023, was how quickly the deposit base could disappear. SVB's vulnerability was not a vulnerability to fraud, or to a single large depositor withdrawing funds. It was a vulnerability to a coordinated withdrawal by a depositor base that was connected to each other in ways that no bank had ever been connected to its depositors before. SVB's depositors were not thirty-seven thousand strangers. They were thirty-seven thousand members of a single industry, who read the same publications, attended the same conferences, sat on the same Slack channels, and took their cues from a small number of venture capital firms whose advice they trusted, in many cases, more than they trusted their own judgment.

The bank's vulnerability, in other words, was not to a bank run in the 1933 sense. It was to a bank run in the 2023 sense — a run that moved at the speed of a forwarded screenshot, in a community that had spent the previous decade learning to make decisions collectively and quickly. The FDIC had been designed to handle the first kind of run. It had not been designed to handle the second.

There were warnings. They were ignored, in the way that regulatory warnings are often ignored — not through conspiracy, but through the slow accretion of bureaucratic process. In the spring of 2022, as the Federal Reserve began raising interest rates, SVB's chief risk officer, Laura Izurieta, left the bank. Her replacement, Kim Olson, was not hired until January 2023, leaving the bank without a formal chief risk officer for much of the year in which its interest rate risk was materializing. The position was, in the organizational chart, the position responsible for identifying and managing precisely the kind of risk that was, during the months it sat vacant, materializing on the bank's balance sheet. SVB's management treated the vacancy as a personnel matter. The Federal Reserve Bank of San Francisco, which supervised SVB, treated it as a concern to be noted in the next examination cycle. The next examination cycle, in the rhythm of bank supervision, was months away.

In May 2022, the Federal Reserve Bank of San Francisco issued a supervisory finding that the bank's interest rate risk management was deficient. The finding was serious — it identified specific weaknesses in the bank's modeling, in its stress testing, and in its governance of the very portfolio that would, ten months later, be the source of the bank's collapse. The finding was communicated to SVB's management. SVB's management prepared a response. The response was reviewed. The review identified further deficiencies. The further deficiencies were communicated. SVB's management prepared a further response. The process continued, in the manner of regulatory processes, throughout the summer and fall of 2022.

In the summer of 2022, the Fed downgraded SVB's supervisory rating — a step that, in principle, should have triggered closer scrutiny. The downgrade moved SVB from a rating that indicated the bank was being managed satisfactorily to a rating that indicated it was not. The downgrade was, in the architecture of bank supervision, a meaningful escalation. It should have prompted the Fed's supervisory team to consider formal enforcement action — a cease-and-desist order, a formal agreement, or a civil money penalty. It did not. The supervisory team continued to work through the

informal channels of examination findings and management responses. The formal channels were available. They were not used.

In October 2022, the bank's internal dashboard showed that its unrealized losses on the held-to-maturity portfolio had grown to the point where they exceeded the bank's tangible equity. The bank was, by its own internal measures, insolvent on a mark-to-market basis — the assets it held, valued at what they would fetch in the market, were worth less than the bank's debts. The dashboard was shared with the Federal Reserve Bank of San Francisco. The Fed's examiners reviewed it. They noted it. They did not escalate the finding to the level of immediate intervention. The reasoning, to the extent it can be reconstructed from the Fed's subsequent post-mortem review, was that the insolvency was "paper" insolvency — the bank could, in theory, hold the securities to maturity and the losses would never be realized. This was technically correct. It was also the same reasoning that had been used to justify inaction in every previous banking crisis in which a bank had been allowed to drift toward failure on the assumption that its losses were temporary.

The Federal Reserve's own post-mortem review of the SVB supervision, published in April 2023, was unusually candid. The review concluded that the Fed's supervisory team had identified the right risks, had communicated them to the bank's management, and had escalated them through the appropriate internal channels — but that the escalation had not produced action proportionate to the risks. The review cited a 2018 change in the law that had raised the threshold for enhanced supervisory scrutiny from fifty billion dollars in assets to two hundred and fifty billion, a change that had exempted SVB from the most stringent post-2008-crisis rules. The review cited a 2019 change in the Fed's own supervisory approach that had, in the name of reducing burden on smaller regional banks, shifted the focus of examination away from the kind of sustained, on-site scrutiny that might have caught the bank's interest rate risk earlier. The review cited, in frank terms, a culture within the supervisory team that had been reluctant to push back against a bank whose management was regarded as sophisticated and whose financial performance, on the metrics the Fed was measuring, had been strong.

None of this was acted upon with the urgency that the situation, in retrospect, demanded. The Fed's supervisory team flagged the issues, but the escalation process — from supervisory findings to formal enforcement action to required remediation — is a process designed to move at the pace of regulatory bureaucracy, not at the pace of a venture capital firm forwarding an email. By the time the Fed's concerns had been formally communicated to SVB's management and the bank had begun, slowly, to take steps to address its interest rate exposure, the situation had become acute.

On February 27, 2023, Greg Becker sold three and a half million dollars of SVB stock. The sale was executed under a Rule 10b5-1 trading plan that Becker had adopted on January 26 — a legal mechanism that allows corporate insiders to sell stock on a predetermined schedule, without being accused of trading on material nonpublic information. The sale was disclosed in an SEC filing. It was entirely aboveboard. It was also the kind of thing that a careful chief executive, whose bank was about to enter a period of acute public scrutiny, might have thought twice about.

On Wednesday, March 8, 2023, SVB announced that it had sold twenty-one billion dollars of its securities portfolio, at a loss of one and eight-tenths billion dollars after taxes, and that it would raise one and three-quarter billion dollars in new equity capital. The announcement was, in the language of

corporate communications, a "strategic action" — an effort to shore up the balance sheet by realizing losses now, while the losses were still manageable, and by raising capital to absorb them. The announcement was made after the market closed, in a press release and on a conference call hosted by Joseph Gentile, the bank's chief administrative officer.

The market's reaction was not what SVB's management had expected. In after-hours trading, SVB's stock dropped sixty percent. Venture capitalists, who had been quietly worried about SVB's balance sheet for months, read the announcement as a confession — not of a problem solved, but of a problem that had become too large to solve quietly. If SVB was selling twenty-one billion dollars of securities at a loss, and raising capital on top of that, the bank's situation was worse than the market had appreciated. The announcement, intended to reassure, had the opposite effect. It was the announcement of a bank that was worried. The market, correctly, became worried too.

* * *

The panic that killed Silicon Valley Bank began, in any operational sense, on the morning of Thursday, March 9. By the time the market opened on the East Coast at nine-thirty, SVB's stock was already in freefall, down another sixty percent in pre-market trading. The after-hours drop on Wednesday evening had been driven primarily by hedge funds and institutional investors who had shorted the stock or sold their positions overnight. The drop on Thursday morning was driven by something different — by the depositors.

The Founders Fund email went out at some point in the early morning, Pacific time. It was a short email, addressed to the founders of the companies in Founders Fund's portfolio, and its substance was simple: in light of SVB's announcement the previous afternoon, Founders Fund was recommending that its portfolio companies withdraw their funds from the bank, ideally by the end of the day. The email did not allege fraud. It did not allege insolvency. It did not say anything especially dramatic. It said, in effect: we don't know if SVB is going to fail, but the cost of withdrawing our money is low and the cost of not withdrawing it, if SVB does fail, is high. We are withdrawing.

This is the logic of every bank run, and it is a logic that, in any individual case, is rational. No single depositor, deciding to withdraw from a bank, is acting irrationally. The depositor does not know whether the bank is solvent. The depositor knows that, if the bank is insolvent, the depositor will lose money. The depositor knows that withdrawing is cheap, and that not withdrawing is potentially catastrophic. The depositor withdraws. The logic is correct. The problem is that, when many depositors follow the logic simultaneously, the bank fails — even if the bank was, in fact, solvent before the withdrawals began. The logic of the individual depositor, multiplied across the depositor base, becomes the cause of the very failure the depositor was trying to avoid.

This has been understood, in the economics literature, since the nineteenth century. The economists Douglas Diamond and Philip Dybvig published a formal model of bank runs in 1983, in which they

showed that bank runs could be self-fulfilling prophecies — that a bank could be fundamentally solvent, and yet be brought down by a panic that the bank's own depositors created. The Diamond-Dybvig model also explained why deposit insurance works: if depositors know that their deposits are insured, they have no reason to panic, and the bank run never begins. Deposit insurance, in the Diamond-Dybvig framework, prevents the panic that causes the failure.

The Diamond-Dybvig model assumes, however, that the depositors are insured. SVB's depositors, overwhelmingly, were not. Ninety-four percent of SVB's deposits were above the FDIC's two-hundred-and-fifty-thousand-dollar insurance limit. The insurance limit, which had been set at two hundred and fifty thousand dollars in 2010, in the aftermath of the 2008 crisis, was designed to cover ordinary retail depositors — individuals with checking and savings accounts. It was not designed to cover the operating accounts of technology startups, which routinely held tens of millions of dollars in cash. A startup that had just raised a fifty-million-dollar Series B and had parked that cash at SVB to make payroll for the next eighteen months had, in the event of an SVB failure, an uninsured exposure of forty-nine million seven hundred and fifty thousand dollars. The FDIC insurance limit, for SVB's depositor base, was a rounding error.

The Founders Fund email was forwarded, throughout the morning of March 9, into the private Slack workspaces of dozens of venture capital firms and founder collectives. By ten o'clock Pacific time, founders who had not received the email were hearing about it from founders who had. By eleven o'clock, the founders who had heard about it were logging into SVB's mobile banking app and attempting to initiate wire transfers. By noon, the app was overloaded. By one o'clock, the bank's wire-transfer system was processing withdrawals at a rate that the system had not been designed to handle. By two o'clock, the rate had roughly doubled.

The withdrawals were, in their mechanics, ordinary. A founder opened the SVB mobile app. The founder selected the account from which to withdraw. The founder entered the amount, the receiving bank's routing number, and the receiving account number. The founder clicked a button to authorize the transfer. The bank, in the ordinary course of business, processed the transfer. There was no teller to consult. There was no branch manager to call. There was no human being at SVB, anywhere in the system, who had the authority to refuse the transfer or to slow it down. The transfers were authorized by the depositors and executed by the bank's systems, automatically, at the speed of the bank's internal network. The founders were withdrawing their money. The bank was, structurally, unable to stop them.

This is the part of the story that the founders of the FDIC, in 1933, could not have anticipated. In 1933, a depositor who wished to withdraw funds from a bank had to physically present himself at the bank's branch, during business hours, and ask a teller to hand over cash or to issue a cashier's check. The teller could refuse, or could delay, or could ask the depositor to come back tomorrow. The bank could, if it chose, simply close its doors and refuse to allow depositors to enter — a step that banks in 1933 frequently took, in the form of "bank holidays," to buy time. The physical architecture of the branch, the teller, and the door was, in effect, a rate-limiting mechanism that prevented withdrawals from happening faster than the bank could process them.

The mobile banking app, however, has no door. It is, in effect, a door that is always open, staffed by no one, and which cannot be closed without shutting down the entire system. SVB could not close it

on March 9 without preventing all of its depositors — including the ones who were not panicking — from accessing their funds, which would have been, in itself, an act of self-destruction. The app stayed running. The money kept leaving.

At some point in the early afternoon, Greg Becker held a conference call with the bank's clients. The call was intended to reassure. Becker told the clients, in substance, that the bank had ample liquidity, that the previous day's announcement had been a strategic action rather than a sign of distress, and that the bank's capital position was strong. He asked the clients to "stay calm" and to "support us the way we have supported you through the years." The call did not have its intended effect. Several of the clients on the call were founders who had already initiated withdrawals that morning. Several more initiated withdrawals during the call. The call was, in retrospect, the moment at which the panic became irreversible — not because of what Becker said, but because the founders on the call could hear, in his voice, the slight strain that comes from a chief executive who is trying very hard to sound calm. They had heard that voice before. They had heard it from their own founders, in the all-hands meetings before the layoffs. They knew what it meant.

By three o'clock Pacific time, the withdrawals had reached a rate that the bank could not sustain. The bank's available cash — the cash it held in its reserve account at the Federal Reserve, plus the cash it could raise on short notice by selling its most liquid securities — was being depleted at a rate that would exhaust it by the end of the business day. The bank's treasury department was, by this point, in continuous communication with the Federal Reserve Bank of San Francisco, which was supervising the bank in real time. The Fed was, in turn, in continuous communication with the FDIC and with the Treasury Department in Washington. The calls between San Francisco and Washington were frequent — every hour, then every half hour, then continuously.

The bank, in a final attempt to slow the bleeding, considered the options available to it under the banking laws. It could borrow from the Federal Home Loan Bank of San Francisco, which is the traditional lender of last resort for California banks in need of short-term liquidity, pledging its mortgage-backed securities as collateral. It did this, in increasing amounts, throughout the afternoon. It could borrow from the Federal Reserve's discount window, pledging the same securities. It explored this. The mechanics of borrowing from the discount window, however, required the bank to pledge collateral that the Fed would accept at a valuation the Fed was willing to credit — and the Fed's valuation of SVB's mortgage-backed securities, given the unrealized losses on the portfolio, was substantially below the face value. The bank could borrow against the securities, but not against enough of them to cover the withdrawals. It could sell securities from its held-to-maturity portfolio, realizing the losses that the bank had spent the previous year pretending did not exist. It began to do this. The realization of the losses, on the scale required to cover the withdrawals, would have wiped out the bank's capital. The bank would, in effect, be admitting to the very insolvency it had been denying.

The bank's options narrowed, through the afternoon, to a single choice: attempt to survive the day, and hope that the panic would subside over the weekend, or accept that the bank could not survive the day and begin to plan for the consequences. The bank's management, by all accounts, attempted the first. The panic did not subside.

By five o'clock Pacific time, the bank was functionally insolvent. The withdrawals had not stopped. The bank had sold what it could sell. It had borrowed what it could borrow. It had exhausted its options. The bank's available cash was, by the end of the business day, insufficient to cover the withdrawals that had already been initiated and that would settle the following morning. The Federal Reserve Bank of San Francisco, in its communication with Washington, conveyed the substance of the situation: SVB could not open for business on Friday morning. The California DFPI, which had legal authority over the bank's charter, began to prepare the order that would close it.

The California Department of Financial Protection and Innovation, which chartered SVB at the state level, closed the bank on the morning of Friday, March 10. The closure was announced at approximately eleven-thirty Pacific time. The FDIC was named receiver. The FDIC's press release, issued shortly after the closure, announced that the bank's insured depositors would have access to their funds by Monday morning, March 13, and that uninsured depositors would receive an "advance dividend" — a partial payment, against their uninsured balances — within the next week. The receivership certificates that the uninsured depositors would receive, the press release explained, would entitle them to a share of whatever the FDIC recovered from the sale of the bank's assets. The recovery, the press release noted, might take years.

The press release was, in the language of bank resolutions, standard. The FDIC had issued similar press releases for hundreds of bank failures over the previous ninety years. The mechanics of the resolution — close the bank on Friday, transfer insured deposits to a healthy bank over the weekend, return insured depositors' funds on Monday, and then spend the next several years winding down the failed bank's assets and distributing the proceeds to uninsured creditors — were the mechanics the agency had been using since the savings and loan crisis of the 1980s. The mechanics assumed that the uninsured depositors would accept their fate, take their receivership certificates, and wait.

The uninsured depositors, in this case, were not interested in waiting. They were, in the main, technology companies that needed their cash to make payroll the following week. A startup with fifty million dollars at SVB, of which only two hundred and fifty thousand was insured, was facing the prospect of being unable to pay its employees on March 15. The startup's CEO was, by Friday afternoon, on the phone with the startup's venture investors, the startup's lawyers, and any member of Congress who would take the call. The venture investors, who had portfolio companies across the technology industry facing the same problem, were on the phone with each other, with the National Venture Capital Association, and with the Treasury Department. The members of Congress, who had technology companies in their districts and who were fielding calls from those companies' founders, were on the phone with the Treasury, with the Fed, and with the FDIC.

The political pressure that built over the weekend of March 11 and 12 was unlike anything the FDIC had faced in its recent history. Ro Khanna, the Democratic congressman whose district included much of Silicon Valley and whose constituents included many of SVB's depositors, publicly called for the government to guarantee all of SVB's deposits. Other members of Congress, from both parties, made similar calls. Eric Swalwell, whose district overlapped with the technology industry's California footprint, called for the same. Patrick McHenry, the Republican chair of the House Financial Services Committee, issued a statement that, while more measured in tone, conveyed the same substantive

demand. The venture capital industry, which had spent the previous decade cultivating an image of indifference to Washington, discovered, over the course of a single weekend, that it cared very much about what Washington thought. The founders who had spent the previous decade building companies that aimed to "disrupt" the existing financial system discovered, over the course of a single weekend, that they wanted very much to be rescued by the existing financial system's deposit insurance fund.

The pressure was not only political. It was operational. By Saturday morning, it had become clear to the Treasury and the Fed that the failure of SVB was producing contagion across the regional banking system. Signature Bank, a New York-based institution with substantial exposure to the cryptocurrency industry, was experiencing a run of its own. First Republic, a San Francisco-based bank with a depositor base similar in profile to SVB's, was showing signs of stress. Western Alliance, Zions, PacWest — regional banks across the American West were seeing their stock prices collapse and their deposit bases come under pressure. The panic that had killed SVB was, in the language of financial crises, spreading. The institutions that had been designed to isolate bank failures — the FDIC's resolution process, the insurance limit, the assumption that one bank's failure would not cause another's — were, in the case of SVB, failing to isolate the failure.

The Treasury, the Fed, and the FDIC spent the weekend in continuous meeting. The meetings were, by all accounts, tense. Janet Yellen, the Treasury Secretary, had publicly stated on Friday that the government was not considering a bailout — a statement that was technically accurate, in the sense that the government was not considering a bailout of SVB's shareholders or bondholders, but that was widely interpreted as meaning that the government would not guarantee uninsured deposits either. The interpretation was wrong, but it had shaped the political expectations for the weekend. Martin Gruenberg, the chairman of the FDIC, was reportedly in favor of a more aggressive intervention. Jerome Powell, the chairman of the Federal Reserve, was, in the manner of central bankers in crisis, focused on the systemic implications. The three officials and their staffs spent the weekend in a series of meetings and phone calls that extended into the early hours of Sunday morning.

The legal question was whether the FDIC had the authority to guarantee uninsured deposits without Congressional authorization. The answer, in the relevant statute, was that the FDIC could do so if the Treasury Secretary, in consultation with the President, determined that doing so was necessary to avoid "systemic risk" — a determination that, in the language of the statute, required the agreement of two-thirds of the FDIC's board and the Fed's board of governors. The systemic risk exception had been used in 2008, to guarantee the deposits of Citigroup and Bank of America, but it had not been used since. The legal threshold was high. The political threshold, given the outcry from the technology industry and the visible contagion in the regional banking system, was arguably higher.

The determination that the situation posed systemic risk was, in substance, a judgment that the failure to guarantee uninsured deposits would produce a cascade of bank runs across the regional banking system that the FDIC's resolution process could not contain. The judgment was, in retrospect, defensible. The uninsured depositors at SVB were not unique — they were, in their financial behavior, indistinguishable from the uninsured depositors at First Republic, at Signature, at PacWest, at every regional bank in the country whose depositor base included businesses with cash balances above the insurance limit. If SVB's uninsured depositors were allowed to lose access to their funds, the uninsured

depositors at every other regional bank would, rationally, conclude that their funds were at risk, and would move them. The result would be a series of bank runs that would, in the absence of a government guarantee, have required the FDIC to resolve multiple regional banks simultaneously — a logistical challenge the agency was not equipped to handle.

The systemic risk exception was invoked on Sunday evening, March 12. The Treasury, the Fed, and the FDIC issued a joint statement announcing that all depositors at SVB — insured and uninsured — would have access to all of their funds on Monday morning. The statement also announced that the FDIC had closed Signature Bank, a New York-based bank with substantial crypto-industry exposure, which had been experiencing a similar run over the weekend. The Deposit Insurance Fund, the FDIC explained, would cover any losses from the guarantee of uninsured deposits, with the funds recovered through a special assessment on the banking industry over the subsequent several years. The taxpayers, the statement emphasized, would not bear the cost. The shareholders and bondholders of SVB and Signature, the statement also emphasized, would not be protected. The management of both banks had been removed. The statement was an effort to draw a line between a "bailout" — which the government was not doing — and a "systemic risk exception" — which it was.

On Monday morning, March 13, depositors at SVB logged into their accounts and found their balances intact. The wires that had been initiated on Thursday and that had not, in the chaos, been processed — were, by and large, processed. The payroll runs that had been scheduled for Wednesday, March 15, were made. The technology industry, which had spent the previous seventy-two hours convinced that it was about to suffer a catastrophe, did not suffer a catastrophe. The system held.

The system held, however, only because the government had, on Sunday evening, decided to abandon the rules that the system had been operating under for the previous ninety years. The FDIC's insurance limit — two hundred and fifty thousand dollars per depositor per bank — had been, since 2010, the explicit boundary of the government's commitment to depositors. The boundary had been crossed, in the case of SVB, not because the FDIC had decided to cross it as a matter of policy, but because the velocity of the run had made the boundary untenable. If the FDIC had adhered to the boundary, the uninsured depositors would have lost access to their funds for an indefinite period, and the political and economic consequences of that loss would have been severe. The FDIC crossed the boundary because the alternative was worse.

The institutions that were designed to manage money — banks, regulators, central banks, clearinghouses, auditors — were designed for a world in which money moved at a speed that the institutions could match. In that world, the institutions could set rules, monitor compliance, and intervene when the rules were broken. In the world we live in now, money moves faster than the institutions can intervene. The rules are still there. The institutions are still there. But the gap between the speed at which money moves and the speed at which the institutions can respond has, in case after case, become the place where the failures occur.

The run on Silicon Valley Bank was not caused by fraud. It was not caused by excessive risk-taking in the conventional sense — the bank's securities portfolio was, by credit standards, conservative. It was not caused by a regulator that was asleep, although the regulator was slow. It was caused by a mismatch between the speed at which a depositor base, alerted by a screenshot in a Slack channel,

could decide to move forty-two billion dollars, and the speed at which an institution designed in 1933 could respond to that decision. The depositors moved their money in six hours. The FDIC's resolution process, which assumes a bank run that takes days or weeks to develop, arrived on Friday morning to find a bank that had been dead since Thursday afternoon.

The FDIC, to its credit, adapted. The systemic risk exception invoked on Sunday evening was a retrofit — an ad hoc modification of the rules, made under pressure, to prevent a bad situation from becoming worse. The retrofit worked, in the narrow sense that SVB's depositors did not lose their money. But a retrofit is, by definition, a response to a failure that has already happened. The institution did not prevent the run, and it did not keep pace with it. It arrived afterward, certified the damage, and changed the rules to absorb the consequences.

The pattern, however, is the thing. The depositors of Silicon Valley Bank, alerted by a screenshot, moved forty-two billion dollars in six hours, and the institution designed to stop them arrived too late. In the chapters that follow, the depositors will be different — Kenyan phone users, Russian banks, Colombian cartels, Indian street vendors, Argentine savers — and the institutions will be different. The pattern will be the same. Money, in the twenty-first century, moves at a speed the institutions built to manage it cannot match. The institutions do not fail because they are stupid. They fail because they were built for a world that no longer exists, and the world has moved on without consulting them.

There is a coda to the SVB story. In the weeks and months after the failure, the FDIC conducted an auction for the remains of the bank — its branches, its loan portfolio, its remaining deposits, and the relationship with the technology industry that had been the bank's defining asset for forty years. The winning bidder was First Citizens Bank, a North Carolina-based institution whose primary claim to fame was its long history of buying the assets of failed banks at favorable prices. First Citizens had bought assets from the failure of First RepublicBank in 1988, from the failure of Guaranty Bank in 2009, and from a string of smaller failures in between. The bank's business model was to wait for other banks to fail, and to acquire their assets at a discount.

The FDIC sold the bulk of SVB's assets to First Citizens at a discount of roughly sixteen and a half billion dollars. The Deposit Insurance Fund absorbed the loss. The fund would, in turn, be replenished by a special assessment on the banking industry — meaning that the cost of SVB's failure would, over the subsequent several years, be borne by every other bank in the United States, in proportion to their deposits. The cost would, eventually, be passed on to bank customers in the form of slightly higher fees and slightly lower deposit rates. The cost would, eventually, be borne by everyone. This is the way the system works. The system works by spreading the cost of failure across the broadest possible base, so that no individual failure produces an individual catastrophe. The system is, in this respect, effective. The system is also, in this respect, a confession that the system cannot prevent failures — it can only socialize them.

In the months after the failure, Greg Becker was replaced by Tim Mayopoulos, a former Fannie Mae executive who had been brought in to manage the bridge bank the FDIC had created to operate SVB's remaining business until a buyer could be found. Mayopoulos, in his first address to the bank's remaining employees, reportedly told them that the bank they had worked for no longer existed, and that the bank they would work for going forward would be a different institution, with different rules,

under different ownership. The employees, by all accounts, listened politely. Many of them, in the following weeks, found jobs at other banks. The technology industry, which had spent forty years banking with SVB, moved its deposits to JPMorgan Chase, to Morgan Stanley, to First Republic (which would itself fail two months later), and to a scattering of smaller institutions. The deposits moved, in many cases, over a single weekend. The mobile banking apps of the receiving banks processed the inflows without notable difficulty.

The FDIC, in the months after the failure, undertook a review of its resolution process. The review's conclusions, when they were published, were measured in tone and pointed in substance. The agency acknowledged that the speed of the SVB run had exceeded the operational capacity of the resolution process. The agency acknowledged that the assumption on which the resolution process had been built — that bank runs would develop slowly enough for the agency to intervene — was no longer reliably true. The agency recommended changes to the process, including faster mechanisms for closing banks, faster mechanisms for transferring insured deposits, and a reconsideration of the insurance limit. The recommendations, at the time of writing, have not been formally adopted. The process by which the recommendations would be adopted, if they were adopted, is the process of regulatory rulemaking — a process that, by design, moves slowly, through notice and comment periods, through interagency review, through Congressional oversight. The process moves at the pace of regulatory bureaucracy. The process does not move at the pace of a venture capital firm forwarding an email.

The chapters that follow are an account of that patchwork — the assemblage of payments rails, offshore dollar issuers, hawala networks, mobile money systems, and parallel currencies that has, over the past twenty-five years, accreted around the formal banking system the way barnacles accrete around a pier. The patchwork was not designed. It emerged, in the manner of most infrastructure, from the accumulated decisions of thousands of people in dozens of countries, most of whom were simply trying to move money from one place to another by the most convenient available route. It works, in the main, rather better than one might expect. It is also, in ways that the following chapters will make apparent, rather more fragile than its daily operation suggests — the kind of fragility that produces no visible cracks, only sudden collapses, of the sort that arrived at Silicon Valley Bank between breakfast and dinner, without warning, and without any institution positioned to stop it.

SVB, in this telling, was less a warning than a rehearsal.

Chapter 2

The Telecom That Became a Central Bank

In the autumn of 2006, a woman in the Kibera slum of Nairobi sat down with a representative of a microfinance institution called Faulu Kenya and repaid a small loan by typing a sequence of digits into a Nokia 1100. The phone was a grey plastic object roughly the size of a bar of soap, with a monochrome screen, an FM radio, a flashlight, and a battery that lasted, on a single charge, the better part of a week. It had been introduced by Nokia in 2003 and had, by 2006, become the most ubiquitous piece of consumer electronics in sub-Saharan Africa, selling for around fifty dollars at kiosks in every market from Lagos to Lusaka. It was not, by any reasonable standard, a sophisticated device. It was, however, the device on which the most successful payments system of the twenty-first century would be built.

The woman in Kibera was participating in a pilot. The pilot was the work of a Vodafone executive named Nick Hughes and a Vodafone engineer named Susie Lonie, who had spent the previous two years designing a system that would allow borrowers of microfinance loans in Kenya to repay those loans via text message. Hughes, who had come to Vodafone's emerging-markets group from a background in African development finance, had noticed something that the development finance community had been noticing for several years: microfinance institutions, which lend small amounts to poor borrowers at relatively high interest rates, spend a substantial fraction of their operational cost on the logistics of collecting repayments. A loan officer must visit each borrower, count the cash, issue a receipt, and carry the cash back to a branch, where it must be counted again and deposited. The cost of this logistics, in the aggregate, consumed a meaningful fraction of the interest the microfinance institution charged. A system that allowed borrowers to repay via text message, the reasoning went, would reduce the cost of microfinance lending and allow the loans to be offered at lower rates, to more borrowers, in more places.

Hughes took the idea to Vodafone's leadership, which agreed to fund a pilot in Kenya — Vodafone held a stake in Safaricom, the Kenyan telecom, and Kenya was a market in which Vodafone had operational presence. The project also attracted funding from an unusual source: the United Kingdom's Department for International Development, or DFID, which had been looking for ways to support the development of financial services in Africa and which saw in Hughes's proposal a possible answer. DFID contributed roughly nine hundred and twenty thousand pounds — a little over a million dollars — as a matching grant, on the condition that the resulting service be made available to the broader public rather than restricted to a single microfinance institution's borrowers. The condition was, in retrospect, the most consequential paragraph in the grant agreement. Without it, M-Pesa might have remained a niche service for microfinance borrowers. With it, M-Pesa was, from the start, a public service.

Lonie, who had been hired to design the user experience, spent much of 2005 and 2006 in Nairobi, working with Safaricom's engineers and with the staff of Faulu Kenya to build a system that would

work on the phones that Kenyans actually owned. The technical challenge was modest, in the sense that the underlying mechanics — sending a text message, crediting an account — were well within the capabilities of the existing mobile network. The user-experience challenge was substantial. The Nokia 1100 had a small monochrome screen, a numeric keypad, and no internet browser. The interface had to work through SMS, with menus navigated by typing numbers. The instructions had to be simple enough for a user who had never used a phone for anything other than calls to follow. Lonie and her colleagues tested the interface with the pilot borrowers, refined it, tested it again, and eventually arrived at a design that worked: the user dialed a short code, entered a PIN, selected "send money," entered the recipient's phone number and the amount, and confirmed. The transaction took under a minute. The recipient received a text message confirming the transfer. The money was available for collection at any M-Pesa agent — typically, the same shop where the user bought airtime.

The pilot, in its narrow purpose, worked. The woman in Kibera repaid her loan. The loan officer did not need to visit her. The cash did not need to be counted or carried. The transaction was recorded, instantly, in Faulu Kenya's books. The premise was validated. Hughes and Lonie and their colleagues at Safaricom prepared to launch the service to the public.

The service launched on March 6, 2007, under the name M-Pesa. The name combined the letter M, for mobile, with the Swahili word *pesa*, meaning money. The launch was modest. Safaricom advertised it on billboards and radio spots. The advertisements emphasized the microfinance use case, which was the use case the designers had built the product for. There was no particular reason to expect that the service would be adopted at a rate that would surprise anyone. The internal projections, which Hughes and Lonie had prepared for Vodafone's board, forecast that M-Pesa would reach perhaps three hundred thousand users in its first year — a respectable number for a niche service, and one that would have justified the investment. The projections, in retrospect, were off by a factor of three.

Within seven months, M-Pesa had one million registered users. Facebook, which had launched three years earlier, had taken roughly twenty months to reach the same milestone. Within two years, M-Pesa had seven million users — roughly a third of Kenya's adult population. Within five years, it had seventeen million users, and was moving, annually, the equivalent of roughly a quarter of Kenya's gross domestic product. By 2023, it was moving the equivalent of more than Kenya's entire GDP, and had become, in the words of one economist who studied it, "the closest thing to a successful central bank that any telecom has ever built."

How M-Pesa grew is the less interesting question. The more interesting one is what it grew into. The service that Nick Hughes and Susie Lonie had designed was a loan repayment system. The service that Kenyans actually used, within months of launch, was something else entirely. They sent money to relatives in villages. They paid for groceries. They settled taxi fares, school fees, and bar tabs. They saved. They paid debts between friends. The microfinance repayment use case, which had been the entire premise of the product, became, within a year, a footnote. The customers had taken the tool the designers had built and used it for the thing they actually needed — a way to move money from one phone to another, instantly, without a bank.

The designers had built a loan repayment system. The customers wanted a payments system. The payments system is what they got.

* * *

To understand why Kenyans adopted M-Pesa so quickly, it helps to understand what moving money in Kenya was like before M-Pesa existed. The year was 2006. The country had a population of roughly thirty-six million people, of whom perhaps a quarter had a bank account. The banking system consisted of a handful of commercial banks — Kenya Commercial Bank, Equity Bank, Standard Chartered Kenya, Barclays Kenya — which operated a combined total of around seven hundred and fifty branches, concentrated in Nairobi, Mombasa, and a few other urban centers. A Kenyan who lived in a rural village, of which there were many thousands, might be a day's travel from the nearest bank branch. The cost of the travel, in bus fare and lost wages, frequently exceeded the cost of the transaction the Kenyan wished to make.

The options available to a Kenyan who wanted to move money from one place to another were, in 2006, limited and unattractive. The most common method, for a person sending cash to a relative in another part of the country, was to give the cash to a bus driver — a man who drove a matatu, the privately operated minibus that served as the principal form of intercity transport in Kenya — and to ask the driver to deliver the cash to a designated recipient at the destination. The driver would, for a small fee, agree to do this. The system worked, in the sense that the cash usually arrived. It also failed, in the sense that the cash sometimes did not arrive — stolen by the driver, stolen by an intermediary, lost in transit. The recipient, in the event of failure, had no recourse. The sender had no receipt. The transaction existed in no ledger. It was, in the language of finance, an unsecured trust-based transfer, intermediated by a bus driver.

The alternative was the post office. The Postal Corporation of Kenya operated a money transfer service called Postapay, which allowed a person to deposit cash at one post office and have it made available for collection at another. Postapay charged a fee of roughly five percent of the amount sent, with a minimum fee of around two hundred Kenyan shillings — about three dollars. The service took, on average, several hours to settle, and required both the sender and the recipient to visit a post office during business hours. The post offices, like the bank branches, were concentrated in urban areas. A rural recipient might be a day's travel from the nearest post office.

The third option was Western Union, which operated in Kenya through a network of agents — typically banks and foreign exchange bureaus. Western Union was reliable and fast, settling transactions within minutes. It was also expensive. A domestic transfer of a hundred dollars through Western Union in Kenya in 2006 cost, in fees, between five and ten dollars — a charge of five to ten percent, on a transaction that might represent a substantial fraction of the sender's monthly income. For international transfers, the fees were higher still. Western Union also required both the sender and the recipient to visit an agent in person, present identification, and complete a form. The system was designed for the diaspora — Kenyans working abroad who wanted to send money home — and was priced and operated accordingly. It was not designed for the ordinary domestic case of a domestic

worker in Nairobi sending twenty dollars to her mother in a village.

The fourth option was a bank wire, which required both parties to have bank accounts. Since three-quarters of Kenyan adults did not have bank accounts, this option was, for most Kenyans, theoretical.

The result of this landscape was that a substantial fraction of the Kenyan economy operated in cash. Wages were paid in cash. Rent was paid in cash. Groceries were bought in cash. School fees were paid in cash, often in installments, carried by hand to the school. The cash economy was inefficient — cash can be stolen, lost, counterfeited, and worn out — but it was the only economy available to most Kenyans. The formal financial system, with its bank accounts and wire transfers and credit cards, existed for the quarter of the population that had bank accounts, and was, for the other three quarters, functionally invisible.

This is the context in which M-Pesa launched. The service did not, at launch, do anything that a bank could not do. It did not offer credit. It did not offer interest on deposits. It did not offer the range of financial products that a bank account offered. What it offered was a single thing — the ability to move money from one phone to another, instantly, for a small fee, without a bank account — and that single thing turned out to be the thing that most Kenyans, in 2007, had been unable to do at any acceptable cost.

The fee structure was modest. A transfer of one thousand Kenyan shillings — about fifteen dollars, and a common transaction size for a domestic remittance — cost the sender thirty shillings, or roughly forty cents. The equivalent transfer through Postapay would have cost around fifty shillings. Through Western Union, around two hundred and fifty. Through a bus driver, nothing in fees, but with a meaningful probability of the money not arriving at all. M-Pesa was cheap, fast, and reliable. It was also available at a network of agents — small shops, typically, that sold airtime and groceries — that sprang up across the country within months of launch. By 2010, there were twenty-two thousand M-Pesa agents in Kenya. By 2023, there were over three hundred thousand. The bank branch network, in the same period, grew from seven hundred and fifty branches to perhaps two thousand. The agent network, by 2010, was already thirty times the size of the branch network. The payments infrastructure of Kenya had been rebuilt, in three years, by a telecom.

The growth was not, in retrospect, mysterious. M-Pesa solved a problem that most Kenyans had, that no existing institution had solved, and that the existing institutions had no particular incentive to solve. The banks served the banked. The post office served the patient. Western Union served the diaspora. The bus drivers served everyone else, at the cost of unreliability. M-Pesa served everyone, cheaply and reliably, through a device — the mobile phone — that a substantial fraction of Kenyans already owned, and that the rest could acquire for fifty dollars.

Why M-Pesa grew, then, is the less interesting question. The more interesting one is why nobody else — the banks, the post office, the central bank — had built it first. The institutions that could have built it had no incentive to do so. The banks earned fees on the existing payments infrastructure, and had no reason to build a cheaper alternative that would cannibalize those fees. The post office was a state-owned enterprise with a monopoly on certain postal services and no particular pressure to innovate. The central bank regulated banks, not telecoms, and had no mandate to build payments

infrastructure. The result was a vacuum — a large unmet need for cheap, fast, reliable payments, in a country where the institutions that might have met the need were either unable or unwilling to do so. M-Pesa filled the vacuum. The telecom, in this case, did not displace the banks. It built, in the empty space where a payments system should have been, a new one.

The reasons the banks had not built M-Pesa themselves illuminate the structural problem that recurs throughout this book. The Kenyan commercial banks in 2006 were not backward. Equity Bank, in particular, had built a substantial business serving lower-income customers — the very segment that M-Pesa would come to dominate — and had pioneered a number of innovations in branchless banking. Kenya Commercial Bank, the largest bank by assets, had a national branch network and served the country's established middle class. The banks had the technology, the customer relationships, and the capital to build a mobile payments system. What they lacked was the incentive.

The economics of banking, in 2006 as in 1926, depend on the spread between the rate a bank pays on deposits and the rate it charges on loans. Payments services — wire transfers, check clearing, foreign exchange — were, for the Kenyan banks, a secondary business, but a profitable one. The fees the banks charged on wire transfers, in particular, were a meaningful source of non-interest income, and the float on the corporate deposits that the banks held — the money that sat in accounts between the time it was deposited and the time it was withdrawn — was a meaningful source of interest income. A mobile payments system that allowed customers to move money instantly, for a small fee, without maintaining a bank account, would have eroded both of these income streams. It would have reduced wire transfer revenue by undercutting the banks' pricing. It would have reduced deposit float by allowing customers to hold money outside the banking system. From the perspective of a bank's chief financial officer, building M-Pesa would have been an act of self-cannibalization.

The telecom, by contrast, had a different economic logic. Safaricom's primary business was selling airtime — the prepaid minutes that Kenyans bought, in small denominations, to make phone calls. Airtime was already, in 2006, sold through a network of small shops across the country, and was already transferred electronically, via scratch cards and, increasingly, via a system called "Sambaza" that allowed one Safaricom customer to send airtime to another. The infrastructure for moving value between phones, in other words, already existed. M-Pesa was, in technical terms, an extension of the airtime transfer system to a different kind of value — cash, denominated in shillings, rather than airtime, denominated in minutes. For Safaricom, the economics were favorable. M-Pesa would generate fee income on transactions. It would increase customer loyalty, since a customer who used M-Pesa was less likely to switch to a competing telecom. And it would generate float — the cash that customers deposited and left in their M-Pesa accounts, which Safaricom held in a pooled account at a commercial bank and on which it earned interest. The float, in particular, became a substantial source of income as M-Pesa grew. By 2013, the interest on the pooled account was generating hundreds of millions of dollars a year for Safaricom — money that, in a traditional banking model, would have accrued to the banks that held the deposits.

The banks, in 2006, did not see this coming. They saw M-Pesa, when it launched, as a curiosity — a small experiment by a telecom in a market that the banks did not consider their own. The banks served customers with bank accounts. M-Pesa served customers without bank accounts. The two

businesses did not overlap, from the banks' perspective. This perspective was, in retrospect, the banks' first mistake. The customers without bank accounts were not, in any fundamental sense, different from the customers with bank accounts. They were the same customers, at a different point in their financial lives. A domestic worker in Nairobi who sent money to her mother in a village was, in 2006, an unbanked customer. The same domestic worker, in 2010, was an M-Pesa customer. The same domestic worker, in 2015, might have been a customer of Equity Bank's mobile banking product, which competed with M-Pesa. The segmentation that the banks had assumed — that there were banked customers and unbanked customers, and that the two groups were distinct — was a fiction that M-Pesa exposed. There were, it turned out, just customers, and they would use whichever service offered the best combination of cost, speed, and accessibility. M-Pesa offered that combination. The banks, for the first decade of M-Pesa's existence, did not.

The banks' second mistake was to assume that the regulatory framework would protect them. The Banking Act defined a bank. The Communications Act defined a telecom. The two categories were separate, and the institutions that fell into each category were regulated by different agencies, under different statutes, with different mandates. The banks assumed that this separation was permanent — that a telecom could not, by law, offer banking services, and that the Central Bank of Kenya would enforce this separation. The assumption was, in the strict letter of the law, correct. M-Pesa did not, technically, offer banking services. It offered payments services, which fell into a regulatory gap between the Banking Act and the Communications Act. The Central Bank declined to close this gap at the moment of launch. By the time the gap was closed — by the National Payments System Act of 2011 — M-Pesa was already moving a quarter of Kenya's GDP, and the closure took the form of a new regulatory category that legitimized M-Pesa rather than shutting it down.

The banks, in petitioning the Central Bank to regulate M-Pesa as a bank, were asking the Central Bank to use the existing framework to shut down a competitor. The Central Bank, in declining to do so, was recognizing that the existing framework could not, in fact, be used for that purpose — M-Pesa did not fit the definition of a bank, and the Central Bank did not have the legal authority to require it to obtain a banking license. The Central Bank's decision was, in this sense, less a policy choice than a recognition of the limits of its own statutory mandate. The Banking Act, written in a world in which banks were the only institutions that moved money, could not contemplate a world in which a telecom moved money. The act had been overtaken by events. The Central Bank, rather than attempting to stretch the act to cover M-Pesa, allowed the act to remain what it was — a law governing banks — and allowed M-Pesa to operate in the space the act did not cover.

* * *

The Central Bank of Kenya, in the spring of 2007, was aware that Safaricom was about to launch a service that would allow Kenyans to move money via text message. The Central Bank's officials had

reviewed the service, in the manner that central banks review things — through meetings, memoranda, and the exchange of formal correspondence. The question the Central Bank had to answer was whether M-Pesa was a banking service, in which case it would fall under the Banking Act and require a banking license, or a telecom service, in which case it would fall under the Communications Act and require only the telecom license that Safaricom already held.

The Banking Act of Kenya, like the banking acts of most countries, defined a bank as an institution that did two specific things: it took deposits, and it lent money. A bank took deposits from the public, with a promise to repay them on demand, and it lent those deposits to borrowers at a higher rate than it paid to depositors, profiting from the spread. The Banking Act required institutions that engaged in this activity to hold a banking license, to maintain minimum capital requirements, to hold reserves at the central bank, and to submit to ongoing supervision by the central bank's examiners.

M-Pesa, as designed, did one of these things but not the other. It took deposits — or, more precisely, it took cash from customers, deposited that cash in a pooled account at a commercial bank, and credited the customers' phones with an equivalent amount of electronic value, called e-float, which could be transferred to other customers or redeemed for cash at agents. The e-float was, in economic substance, a deposit. The customer had given M-Pesa cash, and M-Pesa had given the customer a claim on that cash, payable on demand. But M-Pesa did not lend. The e-float was fully backed, at all times, by cash held in the pooled account. There was no fractional reserve. There was no lending out of deposits. M-Pesa was, in this respect, less a bank than a warehouse — a custodian of cash that issued transferable receipts.

The Communications Act, for its part, regulated Safaricom as a provider of telecommunications services. It did not address the question of whether a telecom could also provide payments services. The Communications Commission of Kenya, which administered the act, had no mandate to regulate money. The Central Bank of Kenya, which administered the Banking Act, had no mandate to regulate telecoms. M-Pesa, in the regulatory architecture of Kenya in 2007, fit into neither category cleanly. It was a payments service offered by a telecom, and the regulatory architecture did not contain a category for "payments service offered by a telecom."

The Central Bank's response to this problem was, in substance, to issue a letter stating that it had no objection to Safaricom launching M-Pesa. The letter, which has become a minor artifact in the literature on mobile money regulation, did not license M-Pesa as a bank. It did not license M-Pesa as a payment service provider. It did not, strictly speaking, license M-Pesa as anything. It simply acknowledged that the Central Bank was aware of M-Pesa, had reviewed its operations, and did not object to its launch. The regulatory stance, translated into plain English, was: we have looked at this thing, we are not sure what it is, and we are going to let it proceed while we figure out whether we should be regulating it.

This stance — sometimes called "regulatory humility," sometimes called "regulatory forbearance," and sometimes, less charitably, called "regulatory helplessness" — was, in retrospect, the most consequential decision in the history of African finance. The Central Bank of Kenya, by choosing not to regulate M-Pesa at the moment of launch, allowed the service to grow. By the time the Central Bank had developed a framework for regulating it, the service had seventeen million users and was moving a quarter of Kenya's GDP. The political and economic cost of shutting it down, or even substantially

modifying it, was, by that point, prohibitive. The Central Bank had, in effect, allowed a private payments system to become a piece of national infrastructure, and had done so by declining, at the critical moment, to decide whether to allow it.

The commercial banks, in the years after M-Pesa's launch, were not pleased. The banks had watched a telecom, in the space of three years, build a payments system that served three times as many Kenyans as the entire banking industry. The banks had also watched M-Pesa siphon off a portion of the fees that the banks had been collecting on wire transfers, foreign exchange, and the deposit float on corporate accounts. The Kenya Bankers Association, the industry's lobbying body, petitioned the Central Bank in 2008 and again in 2010 to investigate M-Pesa, to regulate it as a bank, and to require Safaricom to obtain a banking license. The grounds were reasonable enough. M-Pesa, the banks argued, was taking deposits. The deposits were not insured. The deposits were not subject to reserve requirements. The deposits were not supervised in the manner that bank deposits were supervised. If the pooled account at the commercial bank were ever to fail, or if Safaricom were ever to become insolvent, the seventeen million Kenyans who held e-float would lose their money, and the systemic consequences would be severe.

The Central Bank's response, in substance, was to acknowledge the banks' concerns and to decline to act on them. The response was framed, in the public correspondence, as a measured assessment of the risks and benefits of M-Pesa. The risks were acknowledged — the deposits were, in fact, uninsured, and the failure of Safaricom would, in fact, have been catastrophic. The benefits were also acknowledged — M-Pesa had, in three years, brought payments services to millions of Kenyans who had never had access to them, and had reduced the cost of moving money by an order of magnitude. The benefits, in the Central Bank's judgment, outweighed the risks. The Central Bank did not require Safaricom to obtain a banking license. It did, however, impose certain conditions — the pooled account had to be held at a prudentially regulated commercial bank, the e-float had to be fully backed at all times, and Safaricom had to submit to ongoing reporting requirements. These conditions, which fell short of full banking regulation, were the framework under which M-Pesa operated for the next several years.

The banks' petition, in retrospect, was the moment at which the battle between the old financial system and the new one was joined. The banks had a reasonable case — M-Pesa was, in substance, taking deposits, and the deposits were not regulated in the manner that bank deposits were regulated. The banks also had a less reasonable case — they were losing fee income to a competitor that had built a better product, and they were using the regulatory framework as a tool to slow that competitor down. The Central Bank, in declining to act on the banks' petition, made a judgment that the public interest in cheap, accessible payments outweighed the banks' interest in maintaining their fee structure. The judgment was, by most assessments, the right one. It was also a judgment that the Central Bank would not have been able to make if M-Pesa had been smaller. The reason the Central Bank could decline to regulate M-Pesa was that M-Pesa was already too big to shut down. The regulation followed the adoption. The institution did not lead; it followed.

The institutions that were designed to manage money — banks, regulators, central banks — were designed for a world in which new financial products were introduced slowly, through institutional

channels, and could be assessed, regulated, and either approved or rejected before they reached the public. In that world, the regulator could examine a new product, consult with the industry, draft a rule, publish the rule for comment, finalize the rule, and then allow the product to launch. The process took months, or years. It was designed to take months or years, because the assumption was that the product would not reach significant scale until the regulator had finished its work. M-Pesa broke this assumption. M-Pesa went from launch to seventeen million users in five years, and from launch to a quarter of Kenya's GDP in six. The regulatory process, which was designed to operate on a timescale of years, was operating on a timescale of years. The adoption of M-Pesa, however, was operating on a timescale of months. By the time the regulatory process had produced a framework, the service had already become infrastructure.

The Central Bank of Kenya adapted reasonably well. The framework it eventually put in place — the National Payments System Act of 2011, and the subsequent regulations governing mobile money — was a sensible piece of regulation that recognized mobile money as a distinct category, imposed appropriate safeguards, and allowed the industry to continue growing. The framework was frequently cited as a model for other countries. It was, however, a framework that arrived four years after M-Pesa's launch, by which point the service was already moving a substantial fraction of Kenya's GDP. The framework did not shape the development of M-Pesa. It ratified, after the fact, the development that had already occurred.

The pattern, here, is the same pattern that killed Silicon Valley Bank, translated into a different setting. In the SVB case, the depositors moved faster than the FDIC's resolution process. In the M-Pesa case, the customers adopted faster than the Central Bank's regulatory process. The institution — the FDIC in one case, the Central Bank of Kenya in the other — was designed for a world in which the relevant velocity was slower. The world has changed. The institutions have not.

The consequences of M-Pesa's growth for the daily life of ordinary Kenyans were substantial. They are the consequences that the development economics literature has studied most carefully and that the broader narrative of mobile money has, perhaps, taken too much for granted. A study by the economists William Jack and Tavneet Suri, published in 2014, found that M-Pesa had lifted roughly two percent of Kenyan households — about 194,000 households, or nearly a million people — out of extreme poverty. The mechanism was straightforward. M-Pesa allowed households to receive remittances from relatives working in cities, in the event of a shock — a crop failure, an illness, a job loss. Before M-Pesa, a household hit by a shock had limited options: it could sell assets, borrow from a local moneylender at punitive rates, or rely on the charity of neighbors who were themselves poor. After M-Pesa, a household hit by a shock could, within minutes, receive money from a relative in Nairobi, or from a relative working abroad, and could use that money to buy food, pay medical bills, or keep children in school. The insurance function that formal financial systems provide to the middle classes of developed countries — the ability to draw on a network of credit and savings in the event of a setback — was, for the first time, available to poor Kenyan households. The network was informal. The transfers were small. But the cumulative effect, across millions of households and millions of transfers, was measurable, and it was large.

The effects extended beyond the insurance function. M-Pesa allowed small businesses to receive payments from customers who did not carry cash, and to pay suppliers without traveling to a bank. It allowed farmers to receive payment for produce sold in distant markets, without the intermediation of a broker who might delay payment or abscond with the proceeds. It allowed schools to collect fees without sending a clerk to each village to collect cash. It allowed employers to pay wages without the logistics of cash distribution. Each of these effects, in isolation, was small. In aggregate, they represented a substantial reduction in the friction of economic life — the small costs, in time and money, that accumulate at every point where money changes hands. The reduction in friction, in turn, translated into higher economic activity, more transactions, and, in the studies that have attempted to measure it, measurable improvements in household income and consumption.

The consequences for the banking system were more complicated. The banks, having initially dismissed M-Pesa as a curiosity, came to view it as a threat, and then, gradually, as a partner. Equity Bank launched its own mobile money product, Equitel, in 2014, using a mobile virtual network operator license to compete directly with M-Pesa. Other banks integrated with M-Pesa, allowing customers to move money between their bank accounts and their M-Pesa wallets. The Central Bank, in 2017, mandated full interoperability between M-Pesa and the commercial banks, requiring that customers be able to move money between the systems without friction. The result, by 2020, was a payments landscape in which a Kenyan could move money between a bank account, an M-Pesa wallet, and another bank account, instantly, for a small fee, using a mobile phone. The landscape was, by most measures, the most advanced in Africa, and among the most advanced in the world.

The banks, in this landscape, occupied a different position than they had occupied in 2006. They no longer controlled the payments infrastructure. The payments infrastructure was M-Pesa, and M-Pesa was Safaricom. The banks provided credit, savings products, and the full range of banking services to customers who chose to use them, but the day-to-day business of moving money — the wire transfers, the bill payments, the remittances that had once generated fee income for the banks — had migrated to the telecom. The banks' share of the payments market declined. Their share of the credit market held, and in some segments grew, as mobile money made it easier for banks to assess the creditworthiness of customers whose transaction histories were now visible. The banks, in other words, did not lose their business to M-Pesa. They lost the part of their business that M-Pesa did better, and they retained the part that they did better. The division of labor between banks and telecoms, in Kenya, settled into a new equilibrium — one in which the telecom moved the money, and the bank held the savings and made the loans.

M-Pesa, in the years after the regulatory framework was put in place, continued to grow. It added merchant payments, allowing customers to pay for goods at participating shops by transferring e-float directly to the shop's account. It added bill payment, allowing customers to pay electricity, water, and school fees. It added international remittance, allowing Kenyans abroad to send money home through Vodafone's partner networks. It added, in partnership with the Commercial Bank of Africa, a savings and credit product called M-Shwari, which allowed customers to earn interest on e-float balances and to borrow small amounts against their transaction history. It added, in subsequent years, a fuller suite of financial products — insurance, pension savings, and overdraft facilities — that increasingly resembled

the product suite of a bank. By 2020, M-Pesa was, in functional terms, a bank, a payments processor, a savings institution, and a credit provider, serving thirty million Kenyans and processing transactions equivalent to more than the country's GDP.

The Central Bank of Kenya, throughout this expansion, continued to regulate M-Pesa as something other than a bank — as a mobile money service, with its own regulatory category, its own safeguards, and its own supervisory framework. The category was, in a sense, an admission that the existing categories had failed. The Banking Act had defined a bank. The Communications Act had defined a telecom. Neither definition had contemplated the possibility of a payments service offered by a telecom that would, over the course of a decade, evolve into something that looked, from the customer's perspective, very much like a bank. The Central Bank's solution was to create a new category — mobile money — and to regulate M-Pesa under it. The new category was, like the letter of no objection that had preceded it, a response to a situation that the existing framework could not handle. It was, in the language of regulatory architecture, a patch.

The patch has, in the fifteen years since M-Pesa's launch, been replicated across Africa. Tanzania, Uganda, Ghana, Ivory Coast, Nigeria, and a dozen other countries have, with varying degrees of success, built mobile money systems modeled on M-Pesa, regulated under frameworks that follow the Kenyan template. The replication has not always worked as well as the original — M-Pesa's success was, in retrospect, the product of a specific combination of factors, including a regulator willing to forbear, a telecom with the resources to build the agent network, and a population with a strong need for payments services and no good existing alternatives. The replications have, in some cases, lacked one or more of these factors, and have grown more slowly as a result.

Tanzania, where Vodafone's rival Vodacom launched M-Pesa in 2008, grew more slowly than Kenya — the agent network took longer to build, and the population was more dispersed — but reached comparable penetration by the mid-2010s. Uganda, where MTN launched MTN Mobile Money in 2009, followed a similar trajectory. Ghana, where MTN launched MTN Mobile Money in 2009, faced initial resistance from the Bank of Ghana, which was less inclined than the Central Bank of Kenya had been to forbear, and growth was slower as a result — until the Bank of Ghana, in 2015, relaxed its framework to allow telecoms to operate mobile money without a banking license, at which point growth accelerated dramatically. Nigeria, the largest market in Africa by population, was the slowest of the major markets — the Central Bank of Nigeria, in 2015, issued mobile money licenses to a range of operators but excluded the major telecoms, in the name of preserving the separation between banking and telecoms. The result was a fragmented market in which mobile money grew slowly, and in which, by 2020, Nigeria had one of the lowest mobile money penetration rates in West Africa. The Nigerian decision was, in the literature on mobile money regulation, frequently cited as a cautionary tale — a case in which the regulator's desire to maintain the existing framework prevented the adoption of a service that would have benefited tens of millions of unbanked Nigerians. The Central Bank of Nigeria, in 2021, reversed course and granted mobile money licenses to the major telecoms, and growth subsequently accelerated. The reversal, however, came more than a decade after Kenya's launch, and the cost of the delay — in transactions not made, remittances not sent, and households not lifted out of poverty — is, by any reasonable estimate, substantial.

The contrast between Kenya's experience and Nigeria's is instructive, because it isolates the role of the regulator. The two countries had similar levels of financial exclusion in 2007, similar mobile phone penetration, and similar underlying demand for payments services. Kenya had a regulator that, whether by wisdom or by inaction, allowed M-Pesa to launch and grow. Nigeria had a regulator that, whether by caution or by capture, prevented the telecoms from launching comparable services. The result, a decade later, was that Kenya had the most advanced payments infrastructure in Africa, and Nigeria had one of the least advanced. The difference was not technology. The technology was available in both countries. The difference was the regulatory framework, and the willingness of the regulator to allow a new category of financial service to emerge outside the existing framework.

The template — telecom as payments provider, regulated under a bespoke category that did not exist before the service did — has become the standard for mobile money in the developing world. By 2023, mobile money services operated in over ninety countries, served more than 1.6 billion registered accounts, and processed transactions worth more than a trillion dollars annually. The majority of these services are operated by telecoms, regulated under frameworks that follow, with local variations, the Kenyan model. The frameworks are, without exception, patches — responses to services that had already launched, designed to ratify rather than to shape. In every country that has adopted mobile money, the adoption has preceded the regulation. In no country has the regulation preceded the adoption. The sequence is the same as it was in Kenya: the customers move first, the telecom builds the service, the regulator follows. The institution, in every case, has arrived after the fact.

The developed world, for its part, has not followed this template. In the United States and Europe, payments systems are still, in the main, operated by banks and by the card networks — Visa and Mastercard — that the banks own. The regulatory frameworks in these countries assume that payments are a banking function, and that the institutions providing payments services are either banks or entities licensed as banks. A telecom that wished to launch a payments service in the United States would, in most states, be required to obtain a money transmitter license in each state, to comply with the Bank Secrecy Act's anti-money-laundering requirements, and to submit to supervision by a patchwork of federal and state regulators. The cost of compliance, for a startup, is substantial. For a telecom, it would be prohibitive. The result is that the United States, in 2023, has a payments infrastructure that is, by most measures, slower and more expensive than the one available in Kenya. A Kenyan street vendor can accept a payment of ten shillings — about ten cents — instantly, for free, via a QR code on her phone. An American small business owner pays two and a half percent plus thirty cents — a minimum of thirty cents on a ten-dollar transaction — to accept a payment via Visa or Mastercard. The Kenyan vendor has better payments infrastructure than the American one. The reason is not that Kenya is more technologically advanced. The reason is that Kenya, in 2007, had a regulatory framework permissive enough to allow a telecom to build a payments system, and the United States, in 2023, does not.

The M-Pesa story, then, has two endings, and they point in opposite directions. The first ending is the one told in the development economics literature: a story of financial inclusion, of a telecom that brought payments to millions of people who had been excluded from the formal financial system, of a regulator wise enough to let it happen. The second ending is the one told in the regulatory economics literature: a story of an institution — the Central Bank of Kenya — that could not categorize a new

product fast enough to regulate it before it became infrastructure, and that was forced, by the speed of adoption, to ratify a situation it had not designed. Both endings are true. They are, in fact, the same ending, seen from different angles.

The lesson of M-Pesa is not that mobile money is good or bad, or that the Central Bank of Kenya was wise or foolish. The lesson is that the rate at which a financial product can be adopted, in the age of the mobile phone, has permanently exceeded the rate at which a regulatory framework can be designed, consulted on, and finalized. M-Pesa went from zero to seventeen million users in five years. The Kenyan regulatory framework that eventually governed it took four years to produce. The adoption outran the regulation. The regulation, when it arrived, did not shape the product. It described, after the fact, what the product had already become.

The customers moved first. The institution followed. In following, it did what institutions do — it categorized, it regulated, it imposed a framework. But the framework it imposed was a description of a situation that had already occurred, not a design for a situation that was about to occur. The situation that was about to occur — the one in which a telecom, somewhere in the world, would build a payments system that served tens of millions of people before any regulator had decided what to do about it — was, by the time the Kenyan framework was finalized, already happening in Tanzania, in Uganda, in Ghana, in the Philippines, and in a dozen other countries where regulators had studied the Kenyan model and decided, wisely or otherwise, to follow it.

The institution did not lead, and it did not keep pace. It arrived, in each case, after the customers had already decided what they wanted, and after the telecom had already built it. Its role, in each case, was to ratify the outcome — to issue, in the dignified language of regulation, the equivalent of a letter of no objection.

There is, in the offices of the Central Bank of Kenya in Nairobi, a real-time display that shows the transactions flowing through M-Pesa, second by second, across the country. The display was installed, by some accounts, in 2013, after a visit by a senior Central Bank official to Safaricom's operations center. The official, on seeing the equivalent display at Safaricom, is said to have realized, in the moment, that the Central Bank's authority over M-Pesa was, in practical terms, limited. The Central Bank could issue regulations. It could require reports. It could, in extremis, order Safaricom to change its operations. But it could not, realistically, shut M-Pesa down. The service was too embedded in the daily lives of too many Kenyans. The political cost of shutting it down would have been, in any plausible scenario, fatal to the government that attempted it. The Central Bank, in installing its own display, was acknowledging what had become obvious: it was a spectator to a system it nominally regulated.

The display shows money moving, in real time, to and from every district in Kenya. The amounts are small — a few hundred shillings here, a few thousand there. The aggregate, over the course of a day, runs to hundreds of millions of dollars. Over the course of a year, it exceeds the country's gross domestic product. The screen does not, in any visible respect, resemble a banking operation. It resembles, more nearly, the dashboard of a power grid, or the operations center of a telecom — which is, in fact, what it is. The money moving across it is not, in any technical sense, passing through a bank. It is passing through a payments system built by a telecom, regulated under a category that did not exist

when the system was built, and used, daily, by thirty million people who have come to regard it as the ordinary way that money moves.

The telecom, in other words, became a central bank. The Central Bank of Kenya still issues the shilling, still sets monetary policy, still regulates the commercial banks. But the institution through which thirty million Kenyans move their money, store their money, and measure their money is not the Central Bank. It is Safaricom. The Central Bank regulates the shilling. The telecom runs the rails.

This is the patchwork in its first African iteration. The rails that move the money were built, in seven months, by a telecom, to solve a problem the regulators had not solved. The regulators, when they finally arrived, ratified the rails. By then the rails were carrying the country. They still are.

Chapter 3

Three Guys in the British Virgin Islands

In the summer of 2014, a man named Brock Pierce stood on a stage at a cryptocurrency conference in California and announced that he and two partners had built a new kind of money. Pierce was thirty-three years old, slight, with curly hair and the slightly over-precise diction of a person who has spent his life being looked at. He had been, as a child, an actor — a minor one, but a working one. He had appeared in *The Mighty Ducks*, in 1992, as a member of the eponymous hockey team, and in a handful of other Disney productions in the years that followed. He had left acting in his late teens, had drifted into technology, and had, by the early 2010s, arrived at cryptocurrency, which was at that time a small and somewhat disreputable corner of the technology world populated by libertarians, cipherpunks, and a scattering of entrepreneurs who had been thrown out of more conventional industries for reasons that varied.

Pierce's two partners were Reeve Collins, a marketing executive, and Craig Sellars, a programmer. The three of them had formed a company, originally called Real Coin, that proposed to issue a digital token called Tether. The token would, they said, be backed one-for-one by United States dollars held in a bank account. For every Tether token issued, one dollar would be deposited in the account. For every Tether token redeemed, one dollar would be withdrawn from the account, and the token would be destroyed. The token, in other words, would be a digital representation of a dollar — a stablecoin, in the language that the industry would later adopt — that could be moved around the internet as easily as a Bitcoin, but that would not, unlike Bitcoin, fluctuate in value against the dollar. It would always be worth one dollar, because it would always be backed by one dollar.

The pitch was straightforward, and the technology was unremarkable. The Tether token was issued on top of the Bitcoin blockchain, using a protocol called Omni that allowed users to create custom tokens on top of Bitcoin's underlying ledger. The mechanics of issuing, transferring, and redeeming Tethers were well within the capabilities of any competent programmer. What was novel about Tether was not the technology. What was novel was the promise — the specific, repeated, and (as it would later emerge) not entirely accurate promise that each Tether was backed, one-for-one, by a United States dollar held in a bank account somewhere.

Pierce, Collins, and Sellars incorporated their company in two places. Tether Limited, the operating entity, was incorporated in Hong Kong. Tether Holdings Limited, the parent company, was incorporated in the British Virgin Islands. The British Virgin Islands is a Caribbean territory of the United Kingdom, with a population of roughly thirty thousand people and an economy that consists, in substantial part, of providing corporate registration services to entities that wish to minimize their exposure to taxation, regulation, and public disclosure. The territory hosts, on paper, several hundred thousand incorporated companies — roughly four companies for every resident. Tether Holdings Limited became one of them.

Pierce, Collins, and Sellars did not run Tether for very long. Pierce, whose past included allegations of misconduct dating to his years in the entertainment industry — allegations he has consistently denied — stepped away from the company's public role within a year of its founding, though the extent of his continued involvement remained a matter of some dispute. Collins and Sellars, too, receded from the company's daily operations. By 2015, Tether was being run, in operational terms, by a different group of people — a group connected to Bitfinex, a cryptocurrency exchange that had been founded in 2012 by a French entrepreneur named Raphael Nicolle and that was, by 2015, one of the largest cryptocurrency exchanges in the world.

The Bitfinex connection mattered. Bitfinex and Tether shared executives. Jean-Louis van der Velde, a Belgian businessman who had spent two decades in the technology and finance industries in Asia, became the chief executive of both Bitfinex and Tether. The chief financial officer of both entities was a man named Giancarlo Devasini. Devasini was, by background, a plastic surgeon. He had trained in medicine at the University of Milan, had practiced cosmetic surgery in the city in the early 1990s, and had, through a series of steps that remain somewhat opaque, transitioned into finance and then into cryptocurrency. By the time he arrived at Bitfinex, he had been involved in a number of technology and finance ventures of varying success, and had developed, by the accounts of those who worked with him, a reputation for operational shrewdness and for a willingness to operate in the gray areas where the regulatory architecture had not yet caught up with the technology. By 2015, he was the chief financial officer of the two most important institutions in the cryptocurrency dollar market — Bitfinex, the exchange where most Tethers were traded, and Tether, the issuer of the tokens that were traded there. He was also, by most accounts, the actual architect of Tether's business model — the person who decided what the reserves would be invested in, how the tokens would be issued, and how the company would respond to the regulatory and banking pressures that would, over the next decade, repeatedly threaten to shut it down.

The overlap between Bitfinex and Tether was, for several years, a matter of some delicacy. The two companies shared executives, shared offices, and shared a common corporate group — though they maintained, on paper, separate corporate structures. The reasons for the separation were regulatory. A cryptocurrency exchange that held customer funds was subject to a different set of regulatory requirements than a company that issued dollar-backed tokens, and the maintenance of separate corporate structures allowed each entity to argue that it was subject only to the regulations applicable to its specific function. The separation was, however, more formal than substantive. The executives who ran Bitfinex also ran Tether. The decisions made at Bitfinex about which tokens to list, which markets to support, and which counterparties to trust had direct implications for Tether's business. The decisions made at Tether about how to invest the reserves, how to manage the peg, and how to respond to redemption requests had direct implications for Bitfinex's customers. The two entities were, in operational terms, a single institution. In legal terms, they were two. The gap between the operational reality and the legal form was the gap in which Tether, for the next decade, would operate.

The Paradise Papers leak, in November 2017, made the overlap public. The leak — a trove of 13.4 million documents from offshore law firms, examined by the International Consortium of Investigative Journalists — revealed the extent to which Bitfinex and Tether executives were intertwined, and

confirmed what the cryptocurrency press had long suspected: that the two companies were, in substance, run by the same small group of people. The leak did not, by itself, cause any regulatory action. It did, however, focus public attention on the arrangement, and it provided journalists and regulators with documentary evidence of a corporate structure that the companies had not, until then, publicly acknowledged.

The world's most successful twenty-first-century dollar issuer, in other words, was run by a former child actor, a marketing executive, a programmer, and a plastic surgeon, operating out of Hong Kong and the British Virgin Islands, with a corporate structure that was revealed to the public by a leak of offshore legal documents. The cast was improbable. The business they built was extraordinary.

* * *

To understand what Tether actually is, and why it has proven so difficult for any institution to regulate, it helps to understand the thing it most closely resembles: the Eurodollar market, which is the offshore dollar market that has existed, in one form or another, since the 1950s.

A Eurodollar, despite the name, has nothing to do with the euro. A Eurodollar is simply a United States dollar held outside the United States — deposited in a bank in London, or Zurich, or Singapore, or the Cayman Islands. The dollars are not, physically, in those places; they are, in the modern system, entries on the books of banks. But the legal jurisdiction in which the deposits are held is outside the United States, which means that the deposits are subject to the regulations of the jurisdiction in which they are held, rather than to the regulations of the United States.

The Eurodollar market arose in the late 1950s and early 1960s, for a combination of reasons that included the Soviet Union's desire to hold dollar reserves outside the reach of the American government, the reluctance of American banks to accept large deposits from certain foreign counterparties, and the interest rate differentials that existed between the United States and other markets. By the 1970s, the Eurodollar market had become the largest pool of capital in the world — a multi-trillion-dollar offshore dollar system that operated, in substantial part, outside the direct supervision of any single central bank. The Federal Reserve, which regulated American banks, did not regulate the foreign branches of those banks in the same way. The Bank of England, which regulated banks in London, regulated the London operations of the foreign banks, but not the dollar deposits those banks held. The result was a market that was, in the language of financial regulation, lightly supervised — a place where dollars could be borrowed and lent with a freedom that was not available in the onshore American market.

The Eurodollar market was, for most of its history, a market run by banks. The institutions that held Eurodollar deposits, that made Eurodollar loans, and that created the offshore dollar instruments that circulated in the market were banks — Chase Manhattan, Citibank, Barclays, Deutsche Bank, and a hundred others. The banks were regulated, however lightly, by their home central banks. The deposits

were insured, however imperfectly, by the national deposit insurance schemes of the countries in which the banks were chartered. The system had failures — banks failed, deposits were lost, scandals erupted — but the system was a banking system, subject to banking regulation.

Tether is, in economic substance, an offshore dollar issuer that is not a bank. It takes deposits — in the form of dollars wired to it by customers — and it issues, in return, dollar-denominated liabilities that circulate as a medium of exchange. The liabilities are called Tethers, or USDT, and they trade, on cryptocurrency exchanges and in the broader cryptocurrency market, at a value of one dollar each. The deposits that back the Tethers are held, in the modern operation, in a mix of United States Treasury bills, short-term secured loans, cash, and other liquid assets. The reserves, in other words, are not unlike the reserves that a money-market fund might hold — a portfolio of short-term, high-quality instruments that generate interest income and that can be liquidated to meet redemption requests.

The differences between Tether and a money-market fund are, however, instructive. A money-market fund is registered with the Securities and Exchange Commission, is subject to the Investment Company Act of 1940, is required to publish its portfolio holdings, and is supervised by the SEC's Division of Investment Management. Tether is registered in the British Virgin Islands, is not subject to the Investment Company Act, does not publish its portfolio holdings in real time, and is supervised by no American regulator. A money-market fund is required to maintain a stable net asset value through specific portfolio constraints imposed by the SEC. Tether maintains a stable value through the simple promise that each Tether is redeemable for one dollar — a promise that has, in practice, been honored, but that has never been verified by a full audit.

The differences between Tether and a bank are also instructive. A bank that takes deposits and issues transferable claims on those deposits is subject to banking regulation — capital requirements, reserve requirements, deposit insurance, examination by the relevant supervisor. Tether takes deposits and issues transferable claims on those deposits, but it is not a bank. It is not chartered as a bank in any jurisdiction. It is not subject to capital requirements. It is not subject to reserve requirements. Its deposits are not insured. It is not examined by a banking supervisor. It is, in the legal architecture of financial regulation, a payments company — an entity that facilitates the transfer of value between parties — and it is regulated, to the extent that it is regulated at all, as a payments company rather than as a bank.

This regulatory status is the source of the Tether problem, and it is the problem that the institutions of financial regulation have, for a decade, failed to solve. Tether performs, in functional terms, the same role that a bank performs — it takes deposits, issues transferable claims, and invests the deposits in interest-bearing assets. But Tether is not regulated as a bank, because it is not legally a bank. It is regulated, at most, as a money services business — a category that was designed for check-cashing operations and money transmitters, not for entities that issue a hundred and twenty billion dollars in dollar-denominated liabilities. The regulatory architecture of the United States, and of the broader international system, does not contain a category for what Tether is. The categories that exist — bank, money-market fund, money services business, payment service provider — were each designed for a specific kind of institution, and each carries a specific set of regulatory requirements. Tether does not fit neatly into any of them. It is, in the language of regulatory architecture, an uncategorized entity — a

thing that the system was not designed to handle.

The uncategorized entity has, over the course of a decade, grown into one of the largest dollar institutions in the world. Tether's market capitalization — which is, in economic terms, the total amount of dollar-denominated liabilities it has issued — crossed ten million dollars in 2016, one billion dollars in 2017, twenty billion dollars in 2020, sixty billion dollars in 2022, and one hundred and twenty billion dollars in 2024. The growth was driven, in the main, by the growth of the cryptocurrency market itself. Tether is the principal medium of exchange in cryptocurrency trading. Most cryptocurrency exchanges do not, for regulatory reasons, accept direct dollar deposits from retail customers. Instead, they allow customers to deposit dollars with Tether, receive Tethers in return, and trade those Tethers for Bitcoin, Ethereum, and other cryptocurrencies on the exchange. The Tethers serve, in effect, as a dollar substitute — a way of moving dollars in and out of cryptocurrency exchanges without moving actual dollars through the banking system.

The convenience of this arrangement is obvious. The inconvenience, for the regulators who are supposed to oversee the dollar, is that a hundred and twenty billion dollars in dollar-denominated liabilities has been issued by an entity that is not a bank, not a money-market fund, and not, in any meaningful sense, supervised. The regulators have, over the past decade, made various attempts to bring Tether under their supervision. None of these attempts has succeeded in slowing Tether's growth. The Commodity Futures Trading Commission fined Tether forty-one million dollars in 2021, for misrepresenting the composition of its reserves during the period 2016 to 2018. The New York Attorney General sued Tether and Bitfinex in 2019, alleging that Bitfinex had lost eight hundred and fifty million dollars of customer funds to a shadow bank called Crypto Capital Corporation, and had covered the loss by drawing on Tether's reserves. Tether settled with the New York Attorney General in 2021, paying eighteen and a half million dollars and admitting no wrongdoing. The Department of Justice has, according to various reports, investigated Tether for potential money laundering and sanctions violations, though no charges have been filed.

The fines and the settlements have, in the aggregate, amounted to less than one tenth of one percent of Tether's current market capitalization. The regulatory actions have imposed reporting requirements and have forced Tether to publish more detailed information about its reserves. They have not, however, slowed Tether's growth. Tether's market capitalization has more than doubled since the 2021 settlements. The regulators, in other words, have succeeded in extracting fines and disclosures from Tether. They have not succeeded in controlling it.

* * *

The question of what, exactly, backs a Tether has been the central question of the company's existence since 2017, when the cryptocurrency market's first major boom drove Tether's market capitalization from ten million dollars to over a billion in the space of twelve months. The question is

simple. The answer has, for most of Tether's history, been anything but.

The company's original position, stated on its website and in its marketing materials, was that every Tether was backed one-for-one by a United States dollar held in a bank account. The position was clear, reassuring, and — as it would later emerge — not entirely accurate. The Commodity Futures Trading Commission, in its 2021 settlement order, found that Tether's reserves had, during the period from June 2016 to September 2017, included not only fiat dollars but also commercial paper, secured loans, and other assets, and that at certain points Tether's reserves had been held, in part, by the cryptocurrency exchange Bitfinex — an entity with which Tether shared executives and which was, in any meaningful sense, the same corporate group. The CFTC found, in other words, that Tether had not, during the period in question, been one-for-one backed by fiat dollars, and that the company's statements to the contrary had been misleading.

The discrepancy between Tether's public statements and the actual composition of its reserves was, in retrospect, the predictable consequence of an arrangement in which a small group of people, operating in a lightly regulated jurisdiction, issued dollar-denominated liabilities against a reserve portfolio whose composition they determined themselves, without independent verification. The arrangement was, from the beginning, an invitation to the very conduct the CFTC later identified. The reserves were not verified because no one had the authority to require verification. The composition of the reserves was not disclosed because no one had the authority to require disclosure. The company's statements about the reserves were not checked because no one had the authority to check them. The regulatory architecture did not contain a category for what Tether was, and the regulators who might have asserted authority over Tether — the CFTC, the SEC, the New York Department of Financial Services, the Banking Commissioners of the British Virgin Islands — each operated within narrow jurisdictional limits that did not, individually, cover the full scope of Tether's activities.

Tether's response to the persistent questions about its reserves was, for the first several years of its existence, to publish a series of documents that it described as attestations but that were not, in the technical sense, audits. An audit, in the language of accounting, is a specific kind of examination conducted by an independent certified public accountant, in accordance with established auditing standards, that produces an opinion on whether an entity's financial statements present fairly, in all material respects, its financial position. An attestation, by contrast, is a narrower engagement in which an accountant provides a report on specific subject matter — in Tether's case, the existence of certain bank accounts and the balances in those accounts as of a specific date and time.

The distinction matters, because the two engagements provide different levels of assurance. An audit provides what accountants call "reasonable assurance" — a high, though not absolute, level of confidence that the financial statements are accurate. An attestation provides "limited assurance" — a lower level of confidence, based on less extensive procedures. Tether has, throughout its history, published attestations. It has never published a full audit. The distinction has been the subject of years of criticism from journalists, short sellers, and regulators, and it has been the subject of years of obfuscation from Tether, which has frequently described its attestations in language that suggests they are more comprehensive than they are.

The first attestation Tether published was from a firm called Friedman LLP, in 2017. The Friedman attestation was, even for an attestation, a narrow document. It reported that, as of a specific date — September 15, 2017 — Tether held approximately four hundred and forty-three million dollars in two bank accounts, and that the balances in those accounts exceeded the number of Tethers then outstanding. The attestation did not verify the composition of the reserves over time. It did not verify that the reserves had been continuously backed during the periods between attestations. It did not verify that the bank accounts in question were the only accounts holding Tether's reserves. It was a snapshot, taken on a single day, of two bank balances. Tether presented it as evidence that its reserves were fully backed. Critics pointed out that it proved nothing of the kind.

The Friedman relationship ended, acrimoniously, in early 2018. Tether announced that it had severed ties with the firm, without completing a full audit. The reasons for the severance were not disclosed, though subsequent reporting suggested that Friedman had encountered difficulties in verifying the movement of funds between Tether's accounts and the accounts of related parties, including Bitfinex. Tether, in its public statements, blamed the severance on the complexity of the audit process and the difficulties of working with a conventional accounting firm on an unconventional business. The truth, as so often with Tether, was less clear than the company's statements suggested.

The next attempt at verification came in 2018, when Tether published a legal opinion from a firm called Freeh, Sporkin & Sullivan. The firm was notable for one of its named partners: Louis Freeh, who had served as the Director of the Federal Bureau of Investigation from 1993 to 2001. The presence of a former FBI Director's name on the document lent it an air of official gravitas that the document itself, on close reading, did not entirely support. The Freeh, Sporkin & Sullivan opinion was not an audit, and was not even an attestation in the accounting sense. It was a legal opinion, in which the firm stated that it had reviewed certain documents and that, based on its review, it believed that the funds in Tether's accounts existed. The opinion was carefully hedged. It did not verify the composition of the reserves over time. It did not verify that the reserves were sufficient to cover all outstanding Tethers. It did not, in any operational sense, do what an audit would have done. It was, however, presented by Tether as evidence that its reserves had been verified by a firm associated with a former Director of the FBI. The gap between what the opinion said and what Tether implied it said became, in the cryptocurrency press, a minor scandal — one of many minor scandals that Tether has accumulated over the course of its existence, each of which has been sufficient to generate headlines and insufficient to slow the company's growth.

The pattern that emerged, in the years between 2017 and 2021, was one of escalating regulatory pressure and escalating disclosure, with each escalation revealing more about Tether's reserves while failing to slow the company's growth. The New York Attorney General's 2019 lawsuit, which alleged that Bitfinex had used Tether's reserves to cover an eight hundred and fifty million dollar loss, forced Tether to disclose that its reserves had, at various points, included loans to Bitfinex — a related party — and that the reserves had not, during those periods, been one-for-one backed by fiat dollars.

The eight hundred and fifty million dollar loss, which became known as the Crypto Capital affair, deserves to be described in more detail, because it illustrates the operational reality of Tether's business more vividly than any regulatory filing. Crypto Capital Corporation was a payment processor — a

company that, in the language of the industry, provided banking services to cryptocurrency exchanges that could not obtain banking services from conventional banks. Conventional banks, in the years after 2013, had become increasingly reluctant to provide accounts to cryptocurrency exchanges, citing concerns about money laundering, sanctions compliance, and the reputational risk of associating with the cryptocurrency industry. The cryptocurrency exchanges, which needed bank accounts to process customer deposits and withdrawals, turned to alternative providers — companies like Crypto Capital that were willing to hold customer funds and to process payments on the exchanges' behalf, often through a network of small banks in various jurisdictions.

Bitfinex had used Crypto Capital as its principal payment processor since approximately 2017. By 2018, Bitfinex had transferred roughly eight hundred and fifty million dollars of customer funds to Crypto Capital, for processing and for safeguarding. In late 2018, the funds went missing. Crypto Capital, it emerged, had not been a bank. It had been a shadow payment processor, operating through a web of bank accounts in various jurisdictions, and it had — according to the allegations that later emerged in various legal proceedings — diverted or lost the funds it held on behalf of Bitfinex and other cryptocurrency exchanges. The principal of Crypto Capital, a man named Reginald Fowler, was indicted by the United States Department of Justice in 2019 on charges of bank fraud and money laundering, and was eventually convicted in 2022.

Bitfinex, having lost eight hundred and fifty million dollars of customer funds, faced a liquidity crisis. The exchange could not, in the immediate aftermath of the loss, honor customer withdrawal requests. The solution that Bitfinex's executives arrived at was to draw on Tether's reserves. Tether, which held the dollar deposits that backed its issued Tethers, had substantial cash on hand. Bitfinex, which shared executives with Tether, arranged to borrow from Tether's reserves — a loan of roughly eight hundred and fifty million dollars, secured by Bitfinex's assets, to cover the shortfall created by the Crypto Capital loss. The loan was not disclosed to Tether's customers, who continued to believe that their Tethers were backed one-for-one by dollars held in a bank account. The loan was, in substance, a use of Tether's reserves to bail out an affiliated entity — a use that was, at best, a conflict of interest and, at worst, a fraud on Tether's customers.

The New York Attorney General's lawsuit, filed in April 2019, alleged precisely this. The lawsuit alleged that Bitfinex had used Tether's reserves to cover the Crypto Capital loss, that the use had not been disclosed to Tether's customers, and that the arrangement constituted a fraud on the New York residents who had purchased Tethers. The lawsuit sought to enjoin Bitfinex and Tether from continuing the arrangement, to require the companies to disclose the full extent of the intercompany loans, and to impose civil penalties. The lawsuit did not, however, shut Tether down. It did not freeze Tether's assets. It did not prevent Tether from continuing to issue Tethers. The New York Attorney General's jurisdiction extended to entities that conducted business in New York, and Bitfinex and Tether, while they had New York customers, were not New York entities. The lawsuit proceeded through the New York courts, with various motions and counter-motions, for nearly two years, before settling in February 2021.

The 2021 CFTC settlement formalized the findings about Tether's reserves, imposing a forty-one million dollar fine and requiring Tether to publish quarterly reports on the composition of its reserves.

The quarterly reports, which Tether began publishing in 2021, were produced by a Cayman Islands-based accounting firm called Moore Cayman. They were attestations, not audits, but they were more detailed than the earlier Friedman attestation, and they revealed, for the first time, the specific composition of Tether's reserves.

The composition was, to many observers, surprising. Tether's reserves, as of the first detailed attestation in 2021, included not only cash and Treasury bills but also commercial paper issued by unknown companies, secured loans to unspecified counterparties, and other assets whose liquidity and quality were difficult to assess. The commercial paper holdings, in particular, attracted attention. Commercial paper is short-term unsecured debt issued by corporations. It is, in normal market conditions, a relatively safe investment. In stressed market conditions, it can become illiquid — the holder may be unable to sell it at any price. Tether's commercial paper holdings, at their peak, were estimated at roughly thirty billion dollars, which would have made Tether one of the largest commercial paper holders in the world, larger than many established money-market funds. The identity of the companies that had issued the commercial paper was not disclosed. Tether's critics speculated that some of the issuers might be related parties — companies connected to Tether's executives or to Bitfinex — and that the commercial paper might, in substance, be loans to those related parties disguised as arm's-length investments. Tether denied these allegations. The identity of the commercial paper issuers has never been independently verified.

The 2021 disclosures, rather than undermining confidence in Tether, coincided with a period of accelerated growth. Tether's market capitalization, which had been roughly thirty-five billion dollars at the time of the CFTC settlement in October 2021, grew to eighty billion dollars by the end of 2022, and to one hundred and twenty billion dollars by the end of 2024. The disclosures, in other words, did not slow Tether's growth. They appear, if anything, to have accelerated it, by providing a level of transparency that the market had previously lacked and that the market found sufficient. The market's judgment, however troubling it might seem to a regulator, was that the disclosures were good enough. Tether was not fully audited, but it was attested. Its reserves were not entirely cash, but they were disclosed. The risks were not zero, but they were manageable. The market accepted the risks and moved on.

The most significant shift in Tether's reserves, in the period after the 2021 disclosures, was the migration from commercial paper into United States Treasury bills. The migration began in 2022, in the aftermath of the collapse of Terra/Luna — a different stablecoin project, whose failure in May 2022 wiped out roughly forty billion dollars of value and briefly destabilized the broader cryptocurrency market. Tether, in the days following the Terra/Luna collapse, experienced a brief de-pegging event — its price dropped to roughly ninety-five cents on some exchanges, as holders rushed to redeem their Tethers for dollars. Tether honored the redemptions. Within forty-eight hours, the price had returned to one dollar. Within two weeks, Tether's market capitalization, which had dipped during the panic, had returned to its pre-crisis level. The crisis, in other words, tested Tether's peg and found it intact.

The crisis also, however, focused attention on the composition of Tether's reserves. The commercial paper holdings, which had been a source of market concern since the 2021 disclosures, became a more acute concern in the aftermath of Terra/Luna. The prospect that Tether might be unable

to liquidate its commercial paper holdings in a stress event — and that the illiquidity might, in turn, prevent it from honoring redemptions — became a topic of active discussion among regulators, journalists, and market participants. Tether's response was to eliminate the commercial paper holdings entirely. By the end of 2022, the company announced that its commercial paper holdings had been reduced to zero. The reserves that had been held in commercial paper were moved into United States Treasury bills.

The migration into Treasuries transformed Tether's position in the global financial system in a way that the company's founders could not have anticipated when they issued their first token in 2014. By the end of 2024, Tether held roughly one hundred billion dollars in United States Treasury bills. This made Tether the eighteenth-largest holder of Treasury securities in the world — larger than the central banks of South Korea, Germany, and Mexico, and larger than the sovereign wealth funds of most countries. Tether, the entity incorporated in the British Virgin Islands by a former child actor, a marketing executive, a programmer, and a plastic surgeon, had become one of the most important buyers of United States government debt.

The transformation has implications that the existing regulatory architecture is not equipped to handle. The Treasury Department, which is responsible for issuing the debt that Tether buys, has an interest in Tether's continued solvency — a sudden liquidation of Tether's Treasury holdings, in the event of a crisis, would impose losses on the Treasury market and on the other participants in that market. The Federal Reserve, which is responsible for the stability of the dollar system, has an interest in the stability of an entity that issues a hundred and twenty billion dollars in dollar-denominated liabilities. The Commodity Futures Trading Commission, which has already fined Tether once, has an interest in the accuracy of Tether's disclosures. The New York Department of Financial Services, which has already settled with Tether once, has an interest in Tether's compliance with New York's financial regulations. None of these regulators, however, has full supervisory authority over Tether, because Tether is incorporated in the British Virgin Islands and operates primarily outside the jurisdiction of any single American regulator.

The geopolitical implications of Tether's growth extend beyond the narrow question of who regulates it. Tether has become, in functional terms, an offshore dollar issuer that operates outside the traditional architecture of the international dollar system. The traditional architecture, which has been in place since the Bretton Woods agreement of 1944, is built around central banks — the Federal Reserve, the European Central Bank, the Bank of Japan, and the central banks of the other major economies — that hold dollar reserves, clear dollar payments, and provide dollar liquidity to their domestic banking systems in times of stress. The architecture assumes that dollars, outside the United States, are held and moved by banks, and that banks are supervised by central banks. Tether does not fit this architecture. Tether issues dollars — or, more precisely, dollar-denominated liabilities — to customers who are not banks, in jurisdictions whose central banks do not supervise it, through a network of cryptocurrency exchanges and blockchain transfers that the traditional architecture does not see.

The result is that a substantial and growing fraction of the international dollar economy now operates outside the traditional dollar architecture. Tether is used, by various estimates, to settle

hundreds of billions of dollars in transactions per day — a volume that, if Tether were a country, would make it one of the largest participants in the international dollar system. The transactions are not recorded in the traditional payment systems — they do not pass through SWIFT, they do not pass through the correspondent banking network, they do not pass through the clearinghouses that the Federal Reserve and the other major central banks oversee. They pass through the blockchain, which is a public ledger but which is not, in any operational sense, supervised by any central bank. The Federal Reserve cannot see, in real time, who is moving Tethers, where they are moving them, or why. The Treasury Department's Office of Foreign Assets Control, which is responsible for enforcing sanctions, can investigate specific addresses on the blockchain, but it cannot comprehensively prevent sanctioned parties from using Tether to move dollars.

The use of Tether in jurisdictions subject to American sanctions has, in fact, been a recurring concern of regulators. Tether has been used, according to various reports, to move dollars in and out of Russia, Iran, and Venezuela — jurisdictions where the traditional dollar system has been closed off by sanctions. Tether has also been used, according to blockchain analytics firms, in cross-border payments that bypass the formal banking system — payments that, in some cases, appear to be associated with money laundering, ransomware, and other illicit activity. Tether's response to these concerns has been to cooperate, to a degree, with law enforcement — the company has frozen tokens associated with sanctioned addresses, and has stated that it complies with law enforcement requests. The cooperation has been sufficient to prevent the most aggressive regulatory action, but it has not been sufficient to eliminate the underlying concern, which is that Tether provides a parallel dollar system that operates outside the traditional architecture and that is, by the nature of its design, difficult for any single regulator to control.

The situation that has resulted is one in which a private, offshore, unaudited entity issues dollar-denominated liabilities equivalent to the entire money supply of a medium-sized country, holds a significant fraction of the United States government's debt, and is not subject to the comprehensive supervision of any single regulator. The regulators have, individually, taken what actions they can. The CFTC has fined Tether. The New York Attorney General has settled with Tether. The Treasury Department has, reportedly, investigated Tether's compliance with sanctions regulations. None of these actions has, however, addressed the fundamental problem: the regulatory architecture of the international dollar system does not contain a category for what Tether has become. Tether is, in functional terms, a bank, a money-market fund, and a central bank, all at once. It is none of these things in legal terms. The architecture has no slot in which to put it.

The comparison to the Eurodollar market is, at this point, instructive. The Eurodollar market, in the 1960s and 1970s, posed a similar problem — a large offshore dollar market that operated outside the direct supervision of any single central bank. The problem was, eventually, addressed through a combination of international cooperation, the development of consolidated supervision standards by the Basel Committee on Banking Supervision, and the gradual extension of national regulatory frameworks to cover the foreign operations of domestically chartered banks. The solution took roughly two decades to implement, and it was implemented by regulators who had, at least, the advantage of dealing with institutions that were banks — institutions that fit into an existing regulatory category, even if the

category needed to be extended to cover their offshore operations.

Tether presents a more difficult problem, because Tether is not a bank and does not fit into the existing category. The Eurodollar market was run by banks. Tether is run by a small group of people operating from Hong Kong and the British Virgin Islands. The Eurodollar market's institutions were subject, however imperfectly, to banking regulation. Tether's institution is subject, however imperfectly, to the regulation of the British Virgin Islands' Financial Services Commission — a regulator that is, by most accounts, competent but small, and that is not equipped to supervise an entity that has become one of the largest dollar institutions in the world. The Federal Reserve, which is the ultimate regulator of the dollar, has no direct supervisory authority over Tether, because Tether is not a bank chartered in the United States. The Treasury Department, which is responsible for the integrity of the Treasury market, has no direct supervisory authority over Tether, because Tether is not a primary dealer or a regulated investment fund. The SEC, which is responsible for the regulation of securities markets, has no direct supervisory authority over Tether, because Tether has not been determined to be a security. The CFTC, which has fined Tether, has jurisdiction over commodity derivatives, not over the issuance of dollar-denominated tokens. Each regulator has a piece of the problem. No regulator has the whole problem. The gap between the regulators is where Tether operates.

The pattern, here, is the same pattern that has appeared in the previous two chapters, translated into a different setting. In the Silicon Valley Bank case, the depositors moved faster than the FDIC's resolution process. In the M-Pesa case, the customers adopted faster than the Central Bank of Kenya's regulatory process. In the Tether case, the issuance of dollar-denominated liabilities has grown faster than any regulator's ability to categorize, supervise, and if necessary shut down the issuer. The institution that could, in principle, regulate Tether does not exist. The institutions that do exist — the CFTC, the NYAG, the Treasury, the SEC — each operate within narrow jurisdictional limits that were designed for a world in which dollar issuance was a function of banks, and in which banks were subject to comprehensive supervision by a single regulator. Tether is not a bank, and the world it inhabits is not the world the regulatory architecture was designed for.

The growth continues. Tether's market capitalization, at the time of writing, exceeds one hundred and twenty billion dollars. The company's quarterly profits, derived primarily from the interest income on its Treasury holdings, exceed two billion dollars. Tether is one of the most profitable financial institutions in the world, with a profit margin that would be the envy of any bank and a headcount that would fit comfortably in a single office building. The company's employees, by various estimates, number between one hundred and one hundred and fifty. The Federal Reserve, which is responsible for the stability of the dollar system that Tether operates in, employs roughly twenty thousand. The European Central Bank employs roughly three thousand five hundred. Tether, with a hundred and fifty employees and a hundred and twenty billion dollars in dollar-denominated liabilities, has become one of the most important dollar institutions in the world — and it achieved this status without a banking license, without comprehensive supervision, and without, at any point in its history, publishing a full audit of its reserves.

The regulators, throughout this growth, have done what regulators do. They have investigated, fined, required disclosures, and settled. They have not, however, stopped Tether from growing, because

they have not had the legal authority to stop Tether from growing, and because the political will to grant them that authority has not, so far, materialized. The cryptocurrency industry, which depends on Tether for the dollar liquidity that oils its trading, has lobbied against regulations that would constrain Tether. The Treasury Department, which has an interest in the demand for Treasury securities that Tether's holdings provide, has not pushed aggressively for action that might reduce that demand. The British Virgin Islands, which collects incorporation fees from Tether and has no particular incentive to investigate its largest corporate resident, has not pushed either. The result is a stalemate, in which Tether continues to grow, the regulators continue to watch, and the gap between Tether's functional role and its regulatory status continues to widen.

Tether is the purest expression of the velocity mismatch. The rate at which a private entity can issue dollar-denominated liabilities, in the age of the internet and the blockchain, has permanently exceeded the rate at which the regulatory architecture can categorize, supervise, and if necessary restrain that entity. The institution that could regulate Tether does not exist. The institutions that do exist operate within jurisdictional limits that were designed for a different world. Tether grew up in the gap between those institutions, and it continues to operate in that gap, issuing dollars to whoever wants them, buying Treasury bills with the proceeds, and generating profits that would, in any other context, be the envy of the regulated banking system.

The British Virgin Islands, where Tether Holdings Limited is incorporated, is a territory of roughly thirty thousand people, spread across four dozen islands in the Caribbean Sea. The territory's principal industry, after tourism, is the provision of corporate registration services to entities that wish to operate outside the scrutiny of larger jurisdictions. The territory's Financial Services Commission regulates the entities incorporated there, but its capacity to supervise an entity that has become one of the largest dollar institutions in the world is, by any reasonable assessment, limited. The Commission's staff is small. Its budget is modest. Its expertise in the supervision of cryptocurrency-related businesses is, like the expertise of most regulators, a work in progress. Tether, the largest entity on its books, is an institution whose operations span Hong Kong, the British Virgin Islands, the Cayman Islands, and the United States Treasury market. The Commission regulates the piece of Tether that sits in the British Virgin Islands. The other pieces are regulated, to the extent that they are regulated at all, by other authorities, or by no one.

Tether's founders — the child actor, the marketing executive, the programmer, the plastic surgeon — have, in the decade since they started the company, built something that none of them, in their most ambitious moments, could have anticipated. They built a dollar issuer. They built a Treasury holder. They built, in functional terms, a central bank — an institution that takes deposits, issues currency, invests in government debt, and provides the dollar liquidity that oils a substantial fraction of the cryptocurrency market. They did it without a banking license, without a full audit, and without the comprehensive supervision that any conventional bank would be subject to. They did it in the gap between the regulators, in the space that the regulatory architecture does not cover, in the jurisdiction that was designed for tax optimization rather than for the supervision of systemic financial institutions.

The question of whether Tether is, in any meaningful sense, solvent is one that the market has answered, repeatedly, by continuing to use it. Tether has survived the 2018 Friedman severance, the

2019 Crypto Capital affair, the 2019 New York Attorney General lawsuit, the 2021 CFTC settlement, the 2022 Terra/Luna collapse, and the brief de-pegging event that followed. Each crisis produced predictions of Tether's imminent demise. Each crisis passed. The peg held. The redemptions were honored. The market, having tested Tether's solvency under stress and found it adequate, moved on. The market's judgment may prove, at some future date, to have been wrong. Tether's reserves, which are attested but not audited, may contain assets whose value, in a sufficiently severe stress event, would prove insufficient to cover the liabilities. The risk is real. The risk has, however, been priced — by the millions of customers who continue to hold Tethers, by the cryptocurrency exchanges that continue to list them, and by the broader market that continues to use Tether as the principal dollar substitute in the cryptocurrency economy. The market has decided that the risk is acceptable. The regulators, who might decide otherwise, have not been able to.

The gap is where Tether operates, and the gap is where the next Tether will operate, and the one after that. The regulatory architecture of the international dollar system was designed for a world in which dollars were issued by banks, supervised by central banks, and held in jurisdictions with the capacity and the will to regulate them. That world is gone. The world that has replaced it is one in which dollars can be issued by anyone, anywhere, with a website and a corporate registration in a Caribbean territory, and in which the institutions that are supposed to manage the system have not — and perhaps cannot — catch up. The rate at which a private entity can issue dollar-denominated liabilities, in the twenty-first century, has exceeded the rate at which any institution can stop it. Tether is the proof of that. The next one will be the confirmation, and by the third or the fourth, it will no longer be remarkable.

Chapter 4

The Regulator Who Shot the Messenger

On the morning of June 18, 2020, the board of directors of Wirecard AG, a German payment-processing company that had been, until that morning, one of the most valuable enterprises in Europe, issued a press release containing a sentence that does not appear in the press releases of well-run companies. The sentence said, in substance, that €1.9 billion of cash reported on the company's balance sheet probably did not exist.

The sentence was the culmination of several days of increasingly frantic activity inside Wirecard's Munich headquarters. The company's auditors — Ernst & Young, which had audited Wirecard's accounts for more than a decade — had been unable to confirm the existence of €1.9 billion in cash that Wirecard reported as being held in escrow accounts in the Philippines, on behalf of the company's Asian operations. The auditors had asked to see the money. The company had been unable to show it. The auditors had asked again. The company had suggested that perhaps the auditors were looking in the wrong place. The auditors had checked the places the company suggested. The money was not there either. The Bangko Sentral ng Pilipinas, the Philippine central bank, had been asked to confirm that the funds had entered the Philippine banking system. The central bank had reported, after investigation, that no such funds had entered its system. The €1.9 billion, which had been reported on Wirecard's balance sheet for several years, which had been certified as existing by Ernst & Young in successive annual audits, and which had been relied upon by investors who had purchased Wirecard's stock at prices above €100 a share, was not in the Philippines. It was not anywhere. It had never existed.

The press release was, in the language of corporate communications, an "admission." In the language of the criminal law, it was a confession. Wirecard, which had been included in Germany's DAX 30 stock index since 2006, which had been celebrated by German politicians as a symbol of the country's growing prowess in financial technology, and which had been valued, at its peak, at roughly €24 billion, had been reporting fictitious profits for years. The Asian operations from which a substantial fraction of the company's reported profits derived — the operations that had been held up as evidence of Wirecard's global reach and technological sophistication — were, in substantial part, a fabrication. The "partner companies" in Asia that supposedly processed payments on Wirecard's behalf, and that supposedly held the €1.9 billion in escrow, were entities whose operations no one at Wirecard's Munich headquarters could credibly describe. The profits that Wirecard had reported, year after year, in its audited financial statements, were profits that had been invented.

The man who had presided over this invention was Wirecard's chief executive, Markus Braun. Braun was fifty years old, Austrian, a lawyer by training, and a man who had, over the course of his fifteen years at the head of Wirecard, cultivated a personal style that was unusual for German corporate leadership. He wore the same outfit every day: a black turtleneck, a black blazer, blue jeans, and white sneakers. He spoke, in interviews and at industry conferences, about "digital disruption," about "the cashless society," and about Wirecard's mission to "redefine the way the world pays." The outfit and the

rhetoric were a performance — an effort to present Wirecard, a company that processed payments for online merchants, as a technology company in the mold of the American firms it admired. The performance was, for many years, successful. German investors, who had watched American technology companies achieve valuations that no German company had matched, were eager to find a German equivalent. Wirecard, with its sleek branding and its claims of rapid growth in Asian markets, appeared to be it. The turtleneck was part of the appeal. It said: this is not a traditional German company. This is something new.

The reality, as the press release of June 18 would acknowledge, was rather older. Wirecard's business was a payment-processing operation — a company that handled credit card transactions for online merchants, particularly merchants in high-risk categories (gambling, adult entertainment, dietary supplements) that traditional payment processors were reluctant to serve. The business was real, and it generated real revenue. The problem was that the real revenue was not large enough to justify the valuation that the German stock market had assigned to the company, or the profits that the company had been reporting. The gap between the real business and the reported business was filled, according to the subsequent investigations, by a web of Asian "partner companies" that supposedly processed payments on Wirecard's behalf and that supposedly held cash in escrow for Wirecard's account. The partner companies were real entities, in the sense that they existed on paper. Their operations, and the cash they supposedly held, were not.

Braun was arrested on June 19, the day after the press release, on suspicion of accounting fraud and market manipulation. He was released on bail of €5 million shortly thereafter, and was subsequently re-arrested as the investigation expanded. He has remained in custody since, facing charges that could, if he is convicted, result in a substantial prison sentence. He has denied the charges, and has maintained, through his lawyers, that the fraud was committed by others — principally by Jan Marsalek, Wirecard's chief operating officer, who had been responsible for the company's Asian operations.

Marsalek, as it happened, was not available to confirm or deny this account. On June 22, four days after the press release, Marsalek boarded a private jet at Munich airport, bound for Minsk, the capital of Belarus. He sent a text message to an associate, informing the associate that he was "in a country that doesn't extradite." He has not been seen in public since. Subsequent investigations, conducted by the investigative journalism consortiums Bellingcat and Der Spiegel, and by various Western intelligence agencies, have reported that Marsalek traveled from Minsk to Moscow, and that he has been living, since his flight, under the protection of the Russian state — possibly with the assistance of Russian intelligence services, with which he had reportedly cultivated relationships during his years at Wirecard. The reports, which have not been officially confirmed, describe a man who has undergone plastic surgery, who moves between secured residences in Moscow and its environs, and who is, for all practical purposes, beyond the reach of German law enforcement. The chief operating officer of a DAX 30 company had, over the course of a weekend, transformed into a fugitive living under state protection in a foreign capital. The transformation was unusually dramatic.

* * *

The story of Wirecard's collapse is, in one sense, the story of a fraud. It is, in another sense, the story of a fraud that was visible, for years, to anyone who cared to look, and that was not stopped because the institutions that should have stopped it were looking in the wrong direction. The institutions in question were, principally, two: Ernst & Young, the audit firm that certified Wirecard's accounts, and BaFin, the German federal financial supervisor that was responsible for overseeing the markets in which Wirecard's stock traded. The first institution failed to verify that the money existed. The second institution, when presented with evidence that the money did not exist, chose to investigate the people presenting the evidence.

The audit failure, in its broad outlines, is the easier of the two to understand, though the details remain, at the time of writing, the subject of regulatory investigations and civil litigation. Ernst & Young had audited Wirecard's accounts since 2009. In each of those years, EY issued a "clean" audit opinion — a certification that Wirecard's financial statements presented fairly, in all material respects, the company's financial position. The certification was, in each year, wrong. The financial statements did not present the company's financial position fairly, because the financial statements included €1.9 billion of cash that did not exist, and profits that had been fabricated. EY's audit procedures, which were supposed to include independent verification of the cash balances reported on the balance sheet, did not, in the case of the Philippine escrow accounts, include such verification — or, if they did, the verification was conducted in a manner that failed to detect that the accounts were fictitious. The Financial Reporting Council, the British regulator that oversees auditors, opened an investigation into EY's Wirecard audits in 2020. The investigation, which is ongoing, has focused on whether EY's audit procedures were adequate and whether the firm's partners exercised appropriate professional skepticism.

The audit failure is, however, not the most interesting failure in the Wirecard story. Audit failures happen. Auditors rely on the information that companies provide them, and companies that are determined to deceive their auditors can, for a time, succeed. The more interesting failure — the failure that this chapter is concerned with — is the failure of BaFin, the regulator that was supposed to oversee the market in which Wirecard's stock traded, and that was supposed to protect investors from precisely the kind of fraud that Wirecard committed.

BaFin — the Bundesanstalt für Finanzdienstleistungsaufsicht, or Federal Financial Supervisory Authority — is the German federal agency responsible for the supervision of banks, financial services institutions, insurance companies, and the securities markets. It is headquartered in Bonn and Frankfurt, employs roughly 2,800 people, and operates, like most financial regulators, through a combination of ongoing supervision, examination, and enforcement. Its powers include the ability to investigate market manipulation, to impose fines for violations of securities laws, and to refer cases for criminal prosecution. Its mandate, in the case of a publicly traded company like Wirecard, is to ensure that the market operates fairly — that investors have access to accurate information, that companies disclose

material facts, and that no participant in the market engages in fraud or manipulation.

BaFin's institutional architecture was, at the time of the Wirecard affair, designed primarily for the supervision of banks. Germany's financial system, like the financial systems of most continental European countries, is bank-centered — the principal financial institutions are banks, and the principal regulatory concerns are banking concerns: capital adequacy, liquidity, lending standards, deposit insurance. The securities markets, while important, are secondary to the banking system in the German model, and the regulatory architecture reflects this. BaFin's banking supervision division is its largest and most resourced division. Its securities supervision division, while competent, is smaller and less prominent. The assumption underlying the architecture is that the principal risks to the financial system are banking risks — the risks that a bank will fail, that its depositors will lose money, and that the failure will cascade through the system. The risks posed by non-bank financial institutions — by payment processors, by financial technology companies, by the kind of entity that Wirecard purported to be — are, in the German regulatory architecture, an afterthought.

Wirecard was, legally, a financial services institution, but it was not a bank. It did not take deposits. It did not make loans. It processed payments — a business that, in the German regulatory framework, was subject to a lighter supervisory regime than banking. Wirecard was supervised by BaFin, but it was supervised as a financial services institution, not as a bank. The supervision was, in practice, less intensive than the supervision that a bank of comparable size would have received. BaFin's examiners did not maintain a continuous on-site presence at Wirecard. They did not conduct the kind of detailed, transaction-level examination of Wirecard's operations that they would have conducted at a bank. They reviewed Wirecard's public disclosures, monitored its stock price for signs of manipulation, and relied, for the verification of Wirecard's financial statements, on the audit opinions that EY issued each year.

The political context in which this supervision took place explains it goes some way toward explaining why BaFin, when it became aware of the allegations against Wirecard, was institutionally inclined to view the company as the injured party rather than as the suspect. Germany, in the years before Wirecard's collapse, had been engaged in a quiet but persistent anxiety about its position in the global technology economy. The country had, in the decades after the Second World War, built an industrial economy that was the envy of Europe — automobiles, chemicals, machinery, the products of its engineering tradition. The technology economy of the late twentieth and early twenty-first centuries, however, had largely passed Germany by. The great technology companies of the era — Apple, Microsoft, Google, Amazon, Facebook — were American. The great technology companies of Asia — Alibaba, Tencent, Samsung — were Asian. Germany had SAP, the enterprise software company, and it had a range of mid-sized industrial technology firms, but it did not have a consumer-facing technology champion of the kind that the United States and Asia possessed. German politicians, across the political spectrum, had come to regard this absence as a problem — a sign that Germany, and Europe more broadly, was falling behind in the industries that would define the coming century.

Wirecard, for many German politicians and commentators, appeared to be the answer. The company was German. It was profitable. It was growing rapidly, at a rate that no other German technology company could match. It operated in the payments space — a space that was, in the mid-2010s, at the frontier of financial technology. Its chief executive wore a turtleneck and spoke the

language of digital disruption, which was the language that German politicians associated with the successful technology companies they admired. Wirecard was, for a Germany hungry for a technology champion, precisely what the country wanted. The German finance minister, Olaf Scholz, met with Wirecard's executives. The German chancellor, Angela Merkel, reportedly promoted Wirecard's interests during a state visit to China. The company was, for a time, a darling of the German political establishment — a symbol of the country's potential in the new economy.

The political embrace of Wirecard created, for BaFin, an environment in which the investigation of Wirecard would have been, politically, awkward. This is not to say that BaFin was politically instructed not to investigate Wirecard — there is no evidence of any such instruction, and BaFin has maintained that its decisions were made independently. But the environment in which BaFin operated was one in which Wirecard was regarded, by the political establishment, as a national champion, and in which the allegations against Wirecard were regarded, by the same establishment, as the work of foreign short sellers who were seeking to profit from the destruction of a German success story. The environment was not neutral. It was an environment in which the institutional reflex — the reflex to investigate the short sellers rather than the company — was reinforced by the political context. The reflex was, in this sense, not purely institutional. It was institutional and political, and the two reinforced each other.

The reliance on EY's audit opinions was, in retrospect, the critical vulnerability. BaFin, like most securities regulators, does not independently verify the financial statements of the companies it supervises. It relies on the auditors to do that, and it relies on the auditors' opinions as the basis for its own supervisory judgments. If the auditor's opinion is wrong — if the auditor has certified financial statements that are fictitious — the regulator's supervisory judgments will be wrong too, because they are built on a foundation that has already collapsed. This is what happened with Wirecard. EY certified the accounts. BaFin relied on the certification. The certification was false. The reliance was misplaced. The fraud continued, year after year, because the institution that was supposed to detect it (the auditor) failed to detect it, and the institution that was supposed to supervise the market (the regulator) relied on the auditor's failure.

The more puzzling aspect of BaFin's conduct is not that it failed to detect the fraud — regulators, like auditors, can be deceived — but that it actively investigated the people who were trying to expose the fraud. The people in question were, principally, two groups: short sellers, who had published detailed research reports alleging that Wirecard's accounts were fictitious, and the Financial Times, which had published a series of investigative articles by a journalist named Dan McCrum making similar allegations. Both groups were, as the subsequent investigations would establish, correct. Wirecard's accounts were fictitious. The short sellers and the FT were right. BaFin, faced with the choice of investigating the company that might be committing fraud or investigating the people who were saying the company was committing fraud, chose the latter.

The first major short-seller report was published in April 2015 by a firm called Zatarra Research. The report, which ran to dozens of pages, alleged that Wirecard's Asian operations were fraudulent, that the "partner companies" were shell entities, and that the profits reported from the Asian operations were fabricated. The report was detailed, specific, and accompanied by documentary evidence — contracts, bank statements, and corporate filings from the Asian jurisdictions in question. The report was

published as a PDF on the internet, available to anyone who wished to read it, and was circulated widely among investors and journalists. Wirecard's stock dropped roughly twenty-five percent on the day of publication.

BaFin's response to the Zatarra report was to open an investigation — not into Wirecard, but into whether the short sellers who had published the report had engaged in market manipulation. The investigation was conducted under the framework of the German Securities Trading Act, which prohibits "market manipulation" — a category that includes spreading false or misleading information about a security in order to profit from the resulting price movement. The logic of the investigation was that the short sellers stood to profit from a decline in Wirecard's stock price, and that the report they had published might, therefore, have been an attempt to manipulate the market rather than a good-faith effort to expose fraud.

The logic was not, in itself, unreasonable. Short sellers do profit from price declines, and short-seller reports have, in some cases, been used as tools of market manipulation — publications of false or misleading information designed to drive down a stock's price for the short seller's benefit. The German Securities Trading Act, like the securities laws of most countries, treats this as a form of fraud, and regulators are right to investigate it. The question, in the case of Zatarra and Wirecard, was whether the report was accurate. If the report was accurate — if Wirecard's accounts were, in fact, fictitious — then the short sellers had not engaged in market manipulation. They had engaged in journalism, or in research, and the publication of accurate information, even by someone who profits from the resulting price movement, is not market manipulation. It is the market working as it is supposed to work.

BaFin did not, however, investigate whether the Zatarra report was accurate. It investigated whether the short sellers had manipulated the market. The investigation lasted for months. It produced no findings of manipulation. It was eventually closed, without charges, having consumed the time and resources of BaFin's enforcement staff without producing any result. The investigation did, however, serve a purpose — not for BaFin, but for Wirecard. The company was able to point to the investigation as evidence that the short sellers were under suspicion, and that the allegations in the report were, therefore, not to be trusted. The investigation, in other words, functioned as a form of implicit official endorsement of Wirecard, at the expense of the people who were saying Wirecard was a fraud.

This pattern would repeat, with variations, over the next several years. The Financial Times, beginning in 2016, published a series of investigative articles by Dan McCrum, a journalist based in London who had been covering European financial markets for the paper. McCrum's articles, published primarily on the FT's Alphaville blog, raised questions about Wirecard's Asian operations, about the structure of its partner companies, and about the discrepancy between the profits Wirecard reported and the observable activity of the businesses it claimed to be operating. The articles were carefully sourced, based on public filings, corporate registries, and interviews with former Wirecard employees and industry participants. They were, in their methodology, indistinguishable from the kind of investigative financial journalism that the FT and other major newspapers publish regularly.

BaFin's response, in February 2018, was to open a formal investigation into whether the Financial Times and Dan McCrum had engaged in market manipulation. The investigation was based on the same logic as the Zatarra investigation — that the FT's articles had coincided with declines in

Wirecard's stock price, and that the FT, or parties connected to the FT, might have profited from those declines. The investigation required the FT to produce documents, to answer questions, and to submit to the scrutiny of a foreign regulator that was, in substance, accusing the newspaper of a crime. The investigation lasted for months. It produced no findings of manipulation. It was eventually closed, without charges.

The decision to investigate the FT rather than Wirecard was, at the time it was made, defensible within the narrow logic of BaFin's institutional mandate. BaFin's securities supervision division was responsible for detecting market manipulation, and the pattern — critical articles published, followed by stock price declines — was, on its face, consistent with a manipulation strategy. The investigation was, in this sense, a routine application of the regulator's enforcement tools. What the investigation missed, and what BaFin's institutional architecture did not equip it to see, was that the pattern was also consistent with a different explanation: that the articles were accurate, that the stock price was declining because investors were learning the truth, and that the appropriate regulatory response was to investigate the company whose stock was declining, not the newspaper that was reporting on it.

The failure here was not a failure of individual judgment. It was a failure of institutional design. BaFin was organized to detect market manipulation. It was not organized to detect corporate fraud. The two activities, while related, require different tools, different expertise, and different institutional reflexes. Market manipulation is detected through the analysis of trading patterns — who bought, who sold, when, and in what size. Corporate fraud is detected through the analysis of financial statements — what the company reported, whether the reports are consistent with observable reality, whether the auditors verified the reports. BaFin's securities supervision division had the tools and the expertise for the first task. It did not, as an institution, have the tools and the expertise for the second. When presented with a situation that required the second — a situation in which a company's financial statements appeared to be fictitious — BaFin's institutional reflex was to treat it as a situation that required the first, and to investigate the trading patterns of the people who were saying the statements were fictitious.

The institutional reflex was, in retrospect, the wrong one. At the time, it felt, to the people inside BaFin who were making the decisions, like the right one. They were doing what their training, their procedures, and their institutional culture told them to do. The training, the procedures, and the institutional culture were designed for a world in which the principal threats to market integrity were posed by traders who manipulated prices, not by companies that fabricated profits. The world had changed. The training, the procedures, and the institutional culture had not.

* * *

The KPMG special audit, which was commissioned by BaFin in January 2019 in response to the continuing allegations and to the political pressure that had begun to build in Berlin, was supposed to

resolve the matter. KPMG — one of the Big Four accounting firms, the same category of firm as EY — was engaged to conduct an independent examination of Wirecard's accounts, with particular focus on the Asian operations that had been the subject of the short-seller reports and the FT articles. The audit was, in the language of regulatory practice, a "special audit" — an extraordinary examination, conducted by an independent third party, at the regulator's request, in response to specific concerns.

The special audit took fourteen months. KPMG's staff examined Wirecard's books, interviewed its executives, and attempted to verify the existence and operations of the Asian partner companies. The attempt to verify the Asian operations was, from the beginning, hampered by the same obstacle that had hampered EY's audits and BaFin's supervision: the partner companies were, in their operations, opaque. They were entities incorporated in various Asian jurisdictions, with bank accounts in various banks, and with operations that supposedly consisted of processing payments on Wirecard's behalf. To verify that the operations existed, one would need to examine the bank accounts, the payment flows, and the contracts with the merchants whose payments were supposedly being processed. The bank accounts, it emerged, were in many cases held at banks that were not subject to German supervision, and that were not obligated to provide information to KPMG. The payment flows, to the extent they could be traced, did not appear to support the volume of transactions that Wirecard had reported. The contracts, where they existed, were between Wirecard and entities whose beneficial ownership was, in many cases, unclear.

KPMG's report, published in March 2020, was a carefully hedged document that said, in substance, that the firm could not verify Wirecard's Asian revenues. The report did not say that the revenues were fictitious — KPMG, as an audit firm, was careful to distinguish between "we cannot verify that this is true" and "this is false." The distinction matters in the language of accounting, where the absence of verification is not the same as the presence of falsification. In the language of the market, however, the distinction collapsed. Wirecard's stock, which had been trading above €100 before the report, dropped sharply on its publication. The market read the report as confirmation that the revenues were fictitious, because the only explanation for KPMG's inability to verify the revenues was that there was nothing to verify.

The market was right. The KPMG report was, in retrospect, the beginning of the end for Wirecard, though the end would take another three months to arrive. In those three months, the company's management made a series of increasingly desperate attempts to salvage the situation — to find the €1.9 billion that the auditors could not find, to produce evidence that the Asian operations were real, to convince the markets and the regulators that the allegations were unfounded. The attempts failed. The money could not be found because it did not exist. The operations could not be verified because they were, in substantial part, fabricated. The allegations were not unfounded because they were true.

The collapse, when it came, was swift. On June 18, 2020, the board issued the press release admitting that the €1.9 billion probably did not exist. On June 19, Braun was arrested. On June 22, Marsalek boarded his private jet for Minsk. On June 25, Wirecard filed for insolvency. The company that had been, at its peak, one of the most valuable enterprises in Europe, and that had been included in Germany's premier stock index for fourteen years, was bankrupt in the space of a week.

The aftermath was extraordinary by German standards. The German parliament — the Bundestag — established an investigative committee to examine the failure of BaFin and the audit profession. The committee, which conducted its hearings over the course of 2020 and 2021, heard testimony from BaFin officials, from EY partners, from Wirecard executives, and from the journalists and short sellers who had, for years, been saying what the committee was now investigating. The testimony was, in many cases, uncomfortable. BaFin's president, Felix Hufeld, appeared before the committee and defended his agency's conduct, arguing that BaFin had acted within the limits of its authority and that the failure was, primarily, a failure of the audit profession. The committee was not persuaded. The questioning, by members of parliament from across the political spectrum, was pointed. How had BaFin, which had received detailed allegations from multiple sources over a period of years, failed to investigate the company? Why had it investigated the Financial Times instead? What had possessed the agency to use its enforcement powers against a newspaper that was, as it turned out, publishing accurate reporting? Hufeld's answers — that BaFin had acted within its mandate, that the agency had relied on the auditors, that the investigation of the FT was a routine application of market-manipulation enforcement — were, in their technical correctness, insufficient. The committee wanted to know why the technical correctness had produced a catastrophic outcome. Hufeld did not have a satisfactory answer. He resigned in January 2021, under political pressure that had become, by that point, irresistible. Several other BaFin officials were subsequently investigated for dereliction of duty. The agency's reputation, which had been considerable, was damaged in a way that would take years to repair.

Dan McCrum, the Financial Times journalist whose reporting had been the subject of BaFin's investigation, published a book about the experience in 2022. The book, titled *Money Men*, recounted the years he had spent investigating Wirecard, the pressure he had faced from BaFin's investigation, and the personal and professional toll of being the target of a regulatory enforcement action while trying to publish accurate reporting about a company the regulator was supposed to be supervising. McCrum's account, which is worth reading for anyone interested in the Wirecard affair, is in part a story about journalism — about the patience and the persistence required to investigate a company that is determined to conceal its operations — and in part a story about institutional failure. McCrum had the information. He published the information. The regulator, rather than acting on the information, investigated him for publishing it. The investigation did not stop him from publishing — the FT continued to support his work, and the articles continued to appear — but it did consume his time, his newspaper's resources, and his peace of mind, for a period of years, while the fraud he was exposing continued unabated.

EY, too, faced consequences. The Financial Reporting Council's investigation into the firm's Wirecard audits was expanded, and EY's practices came under scrutiny across Europe. The firm defended its conduct, arguing that it had been deceived by Wirecard's management and that the audit procedures it followed had been appropriate. The defense was, technically, plausible — auditors do rely on the information that companies provide, and sophisticated fraud can be difficult to detect through standard audit procedures. As a broader matter, however, it was unconvincing. EY had audited Wirecard for more than a decade, had collected substantial fees for doing so, and had issued clean

opinions on accounts that were substantially fictitious. The suggestion that the firm bore no responsibility for the failure was, to many observers, difficult to accept.

Marsalek, for his part, remained at large. The German authorities issued an international arrest warrant for him, and Interpol issued a Red Notice — the highest level of international alert. The warrant and the notice were, in practical terms, ineffective. Marsalek was, according to the subsequent investigations by Bellingcat and Der Spiegel, in Russia, and Russia does not extradite its own nationals, or the foreign nationals it chooses to protect, to Germany.

The details of Marsalek's flight, as reconstructed by the investigations, are worth recounting, because they illustrate the degree to which the chief operating officer of a DAX 30 company had, in the years before his departure, constructed a life that was, in retrospect, rather different from the life that one would expect of a German corporate executive. Marsalek was forty years old at the time of his flight. He was Austrian, like Braun, and had joined Wirecard in 2010, rising rapidly to become the company's chief operating officer and the board member responsible for the Asian operations that were, as the subsequent investigations would establish, the locus of the fraud. He was, by the accounts of those who knew him, a man of considerable personal charm and of considerable ambition, with a taste for expensive watches, private aviation, and the company of people connected to intelligence services and to the security apparatus of various countries. His expense reports, which were subsequently examined by the German investigators, reportedly included charges for private falconry expeditions in the Middle East — a detail that captures the man's self-image. He did not see himself as a payments executive. He saw himself as something more interesting.

The investigations by Bellingcat and Der Spiegel, which used flight records, telecommunications data, and interviews with informed sources, reported that Marsalek had traveled from Munich to Minsk on a private jet on June 22, 2020, accompanied by an Austrian security consultant who had reportedly maintained connections to Russian intelligence services. From Minsk, the investigations reported, Marsalek had traveled to Moscow, where he had been met by representatives of the Russian state. The reports suggested that Marsalek had, during his years at Wirecard, cultivated relationships with Russian intelligence — relationships that may have included providing information about Wirecard's clients and about the European payments landscape in exchange for the kind of protection that he would, in June 2020, need. The reports also suggested that Marsalek had undergone plastic surgery to alter his appearance, and that he had been provided with accommodation in secured residences in and around Moscow. The picture that emerged was one of a man who had, over the course of a decade, built the relationships that would allow him to disappear — relationships that were, in their nature, the relationships of an intelligence asset rather than of a corporate executive.

The reports have not been officially confirmed by the Russian or German governments. The Russian government has denied any involvement in Marsalek's escape and has stated that it has no knowledge of his whereabouts. The German government has stated that it has requested Marsalek's extradition, and that the request has been refused. Marsalek's exact whereabouts and status remain, at the time of writing, a matter of informed speculation rather than confirmed fact. What is confirmed is that the chief operating officer of one of Germany's largest publicly traded companies escaped prosecution for a multibillion-euro fraud by boarding a private jet to a country that does not extradite,

and that he has not been seen in public since. The fact that this was possible — that a corporate executive, in twenty-first-century Europe, could evade arrest through a mode of departure more commonly associated with Cold War defectors — is itself a commentary on the limits of the institutional architecture that was supposed to prevent it.

The Marsalek flight reads like a Cold War thriller — a private jet, a text message about non-extradition, a disappearance into a foreign capital, reports of plastic surgery and state protection. The flight is the most cinematic moment in the Wirecard story, and it has been the subject of extensive journalistic and even literary attention. The flight is less interesting, however, than the institutional failure that made it possible. Marsalek fled because he knew, by June 22, that the fraud was about to be exposed, and that his role in it would be discovered. He had known this, in some sense, for years — the short sellers and the FT had been saying it for years — but the knowledge had not, until the final week, carried the urgency that comes from the imminent prospect of arrest. The urgency arrived late because the institutions that were supposed to provide it — the regulator, the auditor, the prosecutor — arrived late. They arrived late because they were looking in the wrong direction, and they were looking in the wrong direction because their institutional architecture told them to. The architecture was designed for a world in which the principal threats to market integrity were posed by traders, not by companies. That world no longer existed, but the architecture had not been updated to reflect the change.

The pattern, here, is the same pattern that has appeared in the previous chapters. In the Silicon Valley Bank case, the depositors moved faster than the FDIC's resolution process. In the M-Pesa case, the customers adopted faster than the Central Bank of Kenya's regulatory process. In the Tether case, the issuance of dollar-denominated liabilities grew faster than any regulator's ability to categorize and supervise the issuer. In the Wirecard case, the distribution of accurate information about a company's financial condition moved faster than the regulator's ability to process that information and act on it. The short-seller reports and the FT articles were published, in full, on the internet, available to anyone who wished to read them. The information moved at the speed of a forwarded PDF. The regulator's response moved at the speed of a bureaucratic investigation. The information outran the response. By the time the response arrived — by the time BaFin had commissioned the KPMG audit, by the time the audit had been completed, by the time the audit's conclusions had been published — the fraud had been ongoing for years, and the people who had committed it had had years to prepare their departures. Marsalek's flight was possible because the regulator arrived too late to prevent it.

The velocity in this case was information velocity — the speed at which accurate information about Wirecard's financial condition could be distributed to the public. That velocity was, by 2015, effectively instantaneous. A short-seller report, once published as a PDF on the internet, was available to every investor, every journalist, and every regulator in the world within minutes. The report could be forwarded, summarized, discussed, and acted upon, all within the same day. The stock price could react, and did react, within minutes of publication. The market, in other words, processed the information at the speed of the internet. The regulator processed the information at the speed of a government agency. The gap between the two speeds was the gap in which Wirecard operated, and the gap in which the fraud continued, for five years after the first detailed allegations were published.

BaFin's institutional architecture was, in this respect, a relic of an earlier era. The architecture was designed for a world in which information about companies moved slowly — through press releases, through annual reports, through the filtered channels of the financial press. In that world, a regulator could investigate allegations at its own pace, because the allegations would not, in the normal course, reach the broader market until the regulator had finished its investigation. The regulator could work behind the scenes, could consult with the company, could gather evidence, and could act, if action was warranted, before the market was aware that there was a problem. The process took months, or years. It was designed to take months or years, because the assumption was that the information would not reach the market until the regulator was ready for it to reach the market.

That world is gone. The short-seller reports and the FT articles about Wirecard reached the market immediately. The market reacted immediately. The investors who read the reports and the articles made their decisions immediately. The regulator, working on its own timescale, arrived after the market had already processed the information and moved on. By the time BaFin had completed its investigations — the investigation into the short sellers, the investigation into the FT, the KPMG special audit — the market had known, for years, what BaFin was investigating. The regulator's investigations were, in a sense, redundant. The market had already done the work. The regulator was confirming, at great expense and over a period of years, what the market had already concluded.

The confirmation was not, however, without value. The market can be wrong. Short sellers can be wrong. Newspapers can be wrong. The fact that the market believes a company is a fraud does not make the company a fraud — the market believed, for years, that Wirecard was a legitimate business, and the market was wrong about that. The regulator's role is to provide the independent verification that the market cannot provide for itself — to examine the evidence, to interview the witnesses, to reach a conclusion that is based on something other than the collective sentiment of investors. The regulator's role, in other words, is to be slow, careful, and authoritative. The problem is that the world in which the regulator operates is fast, noisy, and skeptical of authority. The regulator's slowness, which was once a virtue, has become, in the age of instantaneous information distribution, a liability. The market moves on. The fraud continues. The regulator arrives, eventually, with the correct answer. The answer is, by the time it arrives, of interest primarily to historians.

The Wirecard affair produced, in the years after its collapse, a substantial reform of the German regulatory architecture. BaFin was reorganized, with a new emphasis on the supervision of non-bank financial institutions. The audit profession came under increased scrutiny, with new rules on the rotation of auditors and on the procedures for verifying cash balances. The German parliament passed legislation expanding BaFin's powers, giving it the authority to conduct its own investigations of companies rather than relying on auditors and on market participants. The reforms were, in their specifics, sensible. They addressed the specific failures that the Wirecard affair had exposed. Whether they will address the next failure — the one that has not yet occurred, in a company that has not yet been identified, using a method that has not yet been invented — is a question that the reforms themselves cannot answer. The reforms are designed for the last war. The next war will be different.

The question that the Wirecard affair poses is not whether BaFin should have been faster, or whether EY should have been more skeptical, or whether the German parliament should have passed

the reforms earlier. The question is whether the institutional architecture of financial regulation — the architecture of auditors who certify, regulators who supervise, and prosecutors who enforce — can operate at the speed at which information now moves. The architecture was designed for a world in which information moved slowly, through institutional channels, and could be assessed and acted upon before it reached the broader public. That world is gone. The world that has replaced it is one in which information moves instantly, through channels that the institutions do not control, and in which the institutions must operate, if they operate at all, in a environment that has already processed the information and moved on.

Wirecard was, in this sense, a case study in the limits of the old architecture. The fraud was visible, the information was available, and the market reacted to the information. The institutions that were supposed to protect the market — the auditor, the regulator, the prosecutor — arrived after the market had already reacted. They arrived with the right answer, but they arrived late, and the lateness mattered. The fraud continued for five years after it was first exposed. Billions of euros in investor money were lost, in those five years, to a fraud that the institutions knew, or should have known, was underway. The chief operating officer, who might have provided the most detailed account of how the fraud was committed, escaped to a foreign capital and has not been seen since. The chief executive, who presided over the fraud, sits in a German jail, awaiting a trial that will, when it concludes, provide a narrative of what happened. The narrative will be interesting. It will not bring the money back, and it will not prevent the next fraud. It will, at best, provide a measure of justice — which is what institutions provide, after the fact, when they can no longer provide prevention before it.

Wirecard, in the end, was less a failure of regulation than a failure of the assumptions on which regulation is based. The assumptions — that information moves slowly, that institutions can process it before the market does, that the regulator's slowness is a virtue rather than a liability — were reasonable when they were made. They were made in a world that no longer exists. The world that has replaced them is one in which a short-seller report, published as a PDF on a Tuesday morning, can reach every investor in the world before lunch, can move the stock price before the market closes, and can leave the regulator, working through its procedures at its own pace, arriving at the crime scene after the crime has already been committed, the criminal has already fled, and the witnesses have already gone home.

The regulator, in the Wirecard case, did not fail to investigate. It investigated the wrong target. It investigated the people who were right, because its institutional architecture told it that the people who were right looked, from the regulator's perspective, like the people who were wrong. The short sellers were profiting from Wirecard's decline. The FT was publishing articles that coincided with Wirecard's decline. The pattern — publication, decline, profit — was consistent with manipulation. The architecture said: investigate the pattern. The architecture did not say: investigate whether the publications are accurate, because the architecture assumed that accuracy was the auditor's job, and the auditor had certified that the accounts were accurate. The architecture was a closed loop, and the loop excluded the possibility that the auditor was wrong and the short sellers were right. It excluded that possibility because it was designed for a world in which auditors were reliable and short sellers were not — a world that had, by 2015, ceased to exist.

BaFin, in the years since the Wirecard collapse, has worked to change its architecture. The agency has hired new staff, has reorganized its divisions, and has adopted new procedures for the supervision of non-bank financial institutions. The reforms are, by all accounts, serious. Whether they will be sufficient is a question that will be answered, if it is answered, by the next crisis — which will, in the nature of these things, arrive through a channel the regulator is not watching, at a speed the regulator cannot match, and in a form the regulator's new architecture was not designed to handle. The reforms are designed for the last war. The next one will, as always, be different.

Chapter 5

Forty-Five Minutes

On the evening of July 30, 2012, a team of technology employees at Knight Capital Group's offices in Jersey City, New Jersey, began a software deployment that they had every reason to consider routine. The deployment was an update to a program called SMARS, which was the software Knight used to route orders to the various exchanges on which U.S. equities traded. SMARS was the brain of Knight's market-making operation — the software that decided, in the microseconds between the arrival of a customer order and the execution of that order, where to send the order, at what price, and in what size. The update was intended to add support for a new program that the New York Stock Exchange was launching on August 1, called the Retail Liquidity Program, which would allow market makers to offer slightly better prices to retail investors. Knight, as one of the largest market makers in U.S. equities, needed to be ready.

The deployment proceeded server by server, across the eight server groups that Knight operated. On seven of the eight server groups, the new SMARS code was installed correctly, replacing the old code. On the eighth server group, something went wrong. The precise mechanism of what went wrong has been the subject of some technical dispute, but the substance is clear: the directory on the eighth server group had not been properly cleaned of retired code, and a function called "Power Peg" — a legacy order-routing function that had been decommissioned in 2003, nine years earlier — remained present. The new deployment, in the course of installing the new code, set a flag that reactivated the Power Peg function. The flag was supposed to have been removed. It was not removed. The Power Peg code, dormant for nine years, was now live.

The deployment was completed. The technology team went home. The eighth server group, unknown to anyone at Knight, was now running a piece of software that had been retired before most of the company's current technology staff had been hired. The software would not be triggered until the market opened the next morning, at which point it would begin, with mechanical persistence, to do the thing it had been designed to do — a thing that no one at Knight, in 2012, wanted it to do.

The Power Peg function had a history, and the history is worth recounting, because it illustrates the way that old code accumulates in a system like sediment in a riverbed. Power Peg had been written, in the early 2000s, to handle a specific kind of order routing that had been useful at the time — a function that would aggressively pursue shares at successively better prices, "pegging" the order to the moving market. The function had been used for a period, had been found to have undesirable side effects in certain market conditions, and had been replaced, in 2003, by a more sophisticated routing algorithm. The replacement had involved deploying the new algorithm and instructing the trading desk to stop using Power Peg. It had not involved removing the Power Peg code from the servers. The code had been left in place, on the theory that it would do no harm if it was not called, and on the theory that removing it might introduce new bugs into a system that was working. The theory was not unreasonable. Old code is frequently left in place, in systems of all kinds, on the assumption that

dormant code is harmless code. The assumption is usually correct. It was not correct on the morning of August 1, 2012.

The flag that reactivated Power Peg was a configuration setting — a piece of data that told the SMARS software which routing function to use. The new SMARS code, in the course of its deployment, had set the flag to a value that, in the context of the new code, was correct. In the context of the old code — the Power Peg code that had been left on the eighth server group — the same value meant something different. It meant: activate Power Peg. The new code and the old code disagreed about the meaning of the flag, because the flag had been defined in the era of Power Peg and had been redefined, in the new code, without the old definition being removed. The disagreement was invisible to the deployment process, because the deployment process did not check for disagreements between new code and old code. The deployment process checked that the new code was installed correctly. The new code was installed correctly. The old code was not installed at all — it was simply present, as it had been for nine years, waiting for a flag that would call it back into service.

Knight Capital, at the time of the deployment, was the largest market maker in U.S. equities. The firm handled roughly seventeen percent of the retail order flow in NYSE-listed stocks and roughly twenty percent of the order flow in NASDAQ stocks. When an individual investor clicked "buy" on a retail brokerage website — E*Trade, TD Ameritrade, Scottrade — the order was, in a substantial fraction of cases, routed to Knight, which executed it on one of the exchanges and returned a confirmation to the broker. Knight made its money on the spread — the small difference between the price at which it bought shares from sellers and the price at which it sold shares to buyers. The spread, on any individual trade, was a fraction of a cent per share. The volume, across millions of trades per day, was what made the business profitable. Knight, in 2011, had generated roughly $717 million in revenue and $115 million in net income. The firm had roughly 1,500 employees. Its chief executive, Tom Joyce, was a respected Wall Street veteran who had run Knight since 2002 and who was, in the industry, regarded as one of the more thoughtful and articulate advocates for the market-making business.

Joyce was, by all accounts, a man who understood both the technology and the politics of the market-making business. He had spent his career in the equities divisions of major banks before coming to Knight, and he was fluent in the language of order flow, spreads, and venue competition that defined the industry. He was also a man who trusted his technology team — who believed that the firm's systems were sound, that the firm's procedures were adequate, and that the firm's risk controls were sufficient. The belief was reasonable, in the context of what Joyce knew. The firm had operated for years without a major technology incident. The firm's systems had been reviewed by the SEC and by Knight's own internal auditors. The firm's technology team was experienced and competent. The belief was, in the end, wrong, but it was wrong for reasons that Joyce could not have been expected to anticipate. The reasons were in the code, and the code was not something Joyce had the expertise or the inclination to examine personally. Joyce ran the firm. The technology team ran the code. The division of labor was appropriate. The division of labor was also, on the morning of August 1, the reason that no one in a position of authority knew, in time, what was happening.

Market making, as a business, depends on speed. The spread between the bid and the offer is, in modern markets, so narrow that a market maker can only profit from it if it executes a very large number of trades, very quickly, with very low operational costs. The speed at which a market maker operates is measured in microseconds — millionths of a second. Knight's systems, in 2012, were capable of sending and receiving orders in tens of microseconds, which was competitive but not exceptional. The firm's competitors — firms like Getco, Citadel, and Virtu — operated at similar speeds. The competition among market makers was largely a competition over speed, and over the technology that produced speed: the servers, the network connections, the software, and the proximity to the exchanges' matching engines. Knight's servers were located in data centers in New Jersey, close to the exchanges' systems, because the physical distance between Knight's servers and the exchanges' matching engines affected the time it took for orders to travel — a time measured in microseconds, but a time that mattered.

The speed of the systems created a corresponding challenge for the firm's risk management. A human trader, executing trades by hand, can be stopped by a supervisor who walks over to the desk and says "stop trading." A human trader who is losing money can be identified, by the risk management team, within minutes of the losses beginning, and can be instructed to close out the positions. The human timescale — minutes, hours, the time it takes to walk across a trading floor — is the timescale on which human risk management operates. Knight's systems operated on a different timescale. Knight's systems could execute millions of trades in the time it took a human being to walk from one end of the trading floor to the other. The risk management infrastructure — the dashboards, the alerts, the kill switches — was designed to operate on a human timescale, because the risk management infrastructure was, in the end, operated by humans. The assumption underlying the design was that the humans would be fast enough. The assumption had been tested, in various forms, over the preceding decade, and it had held. It would hold, on the morning of August 1, 2012, for approximately forty-five minutes. Then it would fail.

* * *

To understand what happened in those forty-five minutes, it helps to understand how the U.S. equity market came to be a place where software, rather than human beings, executes the trades. The transformation was recent, and it was disorienting.

For most of the twentieth century, the U.S. equity market was a place where human beings bought and sold shares, on the floors of exchanges, by means of hand signals and shouted orders. The New York Stock Exchange, the principal exchange, was organized around a system of "specialists" — human market makers who stood at posts on the exchange floor and who were obligated to maintain orderly markets in the stocks assigned to them. A customer who wished to buy shares of IBM would call his broker, who would send the order to the exchange floor, where a specialist would match the buy

order with a sell order from another customer, or would buy the shares from his own inventory if no sell order was available. The system was slow — orders took seconds or minutes to execute, and the spreads between bid and offer were wide, because the specialists needed to cover their costs and their risks. The system was also resilient. A specialist who made a mistake could be corrected by a floor supervisor. A market that was moving too fast could be slowed by a trading halt. The human beings on the floor provided a rate-limiting mechanism that prevented the market from doing things faster than the people running it could understand.

The specialist system had its virtues, and it had its vices, and the vices were, by the 1990s, becoming difficult to ignore. The specialists, as human beings with discretion over the matching of orders, were in a position to profit from the information they possessed about the order flow — to buy before they knew prices would rise, and to sell before they knew prices would fall. The practice, known as "front-running," was illegal, but it was difficult to detect and difficult to prove, and the specialists, who were the only ones with a complete view of the order book, had opportunities that other market participants did not. The specialists were also, as a practical matter, a bottleneck. The volume of trading that a human specialist could handle was limited by the speed of human cognition and human speech. A specialist who was matching orders by shouting across a trading floor could handle, perhaps, a few hundred orders per minute. The market, in the 1990s, was generating orders at a rate that exceeded this capacity, and the excess was being absorbed by the specialists' capital, which was finite. The system was, in the language of engineering, reaching its throughput limit.

The transformation began in the 1970s, with the introduction of electronic trading systems, and accelerated in the 1990s and 2000s, with the rise of alternative trading venues, the decimalization of stock prices (which narrowed spreads and made high-volume trading more profitable), and the emergence of firms like Knight that used technology to execute trades at speeds and volumes that human specialists could not match.

Decimalization, which occurred in 2001, is the change that made the modern market-making business possible. Before decimalization, stock prices were quoted in fractions — sixteenths, eighths, quarters — with the minimum increment being one-sixteenth of a dollar, or 6.25 cents. The minimum spread between the bid and the offer was, therefore, 6.25 cents. A market maker who bought at the bid and sold at the offer made 6.25 cents per share, on every trade, regardless of the volume. The wide spreads made market making profitable at low volumes, and they made the specialist system viable — a specialist could make a living on a few hundred trades per day, because each trade generated a substantial spread. Decimalization reduced the minimum increment to one cent, and the minimum spread to one cent. The spreads collapsed. A market maker who had been making 6.25 cents per share was now making one cent per share, or less. The profitability of market making, at the volumes that had prevailed under the fractional system, was destroyed. The market makers who survived decimalization were the ones who could make up in volume what they had lost in spread — the ones who could execute tens of thousands of trades per day instead of hundreds, and who could do so at low cost. These were the firms that had invested in technology — in the servers, the network connections, and the software that could route orders at machine speed. Knight was one of these firms. The specialists, who had stood on the exchange floor, were not.

The New York Stock Exchange, which had been the dominant venue for U.S. equity trading for most of the twentieth century, saw its market share erode as electronic exchanges — NASDAQ, BATS, Direct Edge, and others — attracted order flow with faster execution and lower costs. The specialists, who had stood at the posts on the exchange floor, were replaced, gradually and then suddenly, by computers in data centers in New Jersey. The New York Stock Exchange itself, which had been a mutual organization owned by its members, converted to a for-profit corporation in 2006 and merged with Euronext in 2007. The floor of the exchange, which had been the center of American capitalism for two hundred years, became a television studio — a place where the opening bell was rung for the cameras, and where a diminishing number of human beings executed a diminishing fraction of the market's trades.

By 2012, the U.S. equity market was a market in which the vast majority of trades were executed by computers, at speeds measured in microseconds, across a dozen competing exchanges and a larger number of alternative trading venues. The human beings who had once stood on the exchange floor were gone, or nearly gone. The market was now run by software — by the smart order routers that decided where to send orders, by the matching engines that paired buyers with sellers, and by the risk management systems that monitored the positions and the exposures. The software was written by human beings, tested by human beings, and supervised by human beings, but the execution was done by the software, at a speed that human beings could not match.

The speed of execution created a corresponding need for speed in risk management. A firm that executes trades in microseconds must monitor its positions in microseconds, and must be able to stop trading in microseconds, because a malfunctioning algorithm can lose more money in one second than a human trader can lose in a day. The industry's response to this need was the development of automated risk controls — pre-trade limits that would reject orders exceeding specified parameters, post-trade monitors that would alert risk managers to unusual position accumulation, and kill switches that would halt trading on a firm's systems when the monitors detected a problem. The risk controls were, like the trading systems themselves, software — programs written to detect the conditions that the trading programs might produce, and to intervene when those conditions arose.

The risk controls were, however, only as good as the assumptions on which they were based. A pre-trade limit that rejects orders exceeding a specified size will stop a single oversized order, but it will not stop a malfunctioning algorithm that sends millions of small orders, each within the limit, that collectively accumulate an enormous position. A post-trade monitor that alerts risk managers to unusual position accumulation will detect the accumulation, but only after the accumulation has occurred — and if the algorithm is fast enough, the accumulation may be catastrophic by the time the alert sounds. A kill switch that halts trading on a firm's systems will stop the bleeding, but only if the kill switch is designed to halt the specific system that is malfunctioning, and only if the humans who operate the kill switch can identify, in real time, which system is malfunctioning and which kill switch to pull.

The Knight Capital incident exposed the limitations of all three types of risk control. The pre-trade limits did not catch the Power Peg orders, because each individual order was within the specified parameters. The post-trade monitors detected the position accumulation, but the accumulation was so rapid — hundreds of millions of dollars in minutes — that the alerts arrived too late to prevent the

losses. The kill switch, when it was finally located and activated, was not designed to stop a single server group from running rogue code while the rest of the firm's systems operated normally. The risk controls were designed for a world in which the trading systems operated at the speed of the risk controls. The trading systems operated at a speed that the risk controls could not match.

The regulatory framework that governed this market was, like the risk controls, designed for an earlier era. The Securities and Exchange Commission, which regulates the U.S. securities markets, had approved Knight's systems through a process that examined the firm's technology, its risk controls, and its supervisory procedures. The approval was based on the assumption that Knight's systems would operate as designed, and that the risk controls would function as intended. The SEC did not, in its approval process, attempt to verify that every line of Knight's code was correct, or that every deployment procedure was followed without error. The SEC's examiners reviewed Knight's policies, examined its systems at a high level, and relied on Knight's representations that its technology was sound. The reliance was reasonable. The framework assumed that the firms it regulated would operate competently, and that the firms' internal controls would catch errors before they became catastrophic. The framework was designed for a world in which errors were caught by humans, in time to prevent disaster. The world had changed. The framework had not.

The New York Stock Exchange's Retail Liquidity Program, the launch of which had prompted Knight's software update, was itself a product of the competitive dynamics of the modern market. The NYSE, which had been losing market share to electronic competitors for years, had designed the Retail Liquidity Program as a way to attract retail order flow back to the exchange. The program allowed market makers to offer retail investors prices that were slightly better than the public quotes — a fraction of a cent per share better, which was enough to be attractive to retail brokers who routed orders on behalf of their customers. The program required market makers to update their systems to support the new order types, and it required them to do so by August 1, the launch date. Knight, like every other market maker that wished to participate, had to update its systems by the deadline. The deadline created pressure. The pressure led to a deployment that was, by all accounts, performed with appropriate care — but that was performed under time constraints, and that contained an error that the care did not catch.

* * *

The market opened at nine-thirty in the morning Eastern time on August 1, 2012. The opening of the market is, in normal circumstances, an unremarkable event — the matching engines begin accepting orders, the market makers begin quoting, and trading commences at a volume and pace that build over the first few minutes as the various participants ramp up their systems. On the morning of August 1, the opening was, for the first few seconds, unremarkable. Then the eighth server group began to receive order flow.

The Power Peg function, reactivated by the flag that the deployment had failed to remove, began to execute. The function had been designed, in the early 2000s, to route orders in a specific manner that had been useful at the time but that had been superseded by subsequent improvements in Knight's order-routing logic. The function had been decommissioned in 2003, and had been left in the codebase rather than being removed — a common practice in software development, where old code is often retained on the assumption that it will do no harm if it is not called. The flag that the new deployment set, however, called it. The function began to do what it had been designed to do: it routed orders. The routing it performed, however, was not the routing that the new SMARS code was supposed to perform. The Power Peg function routed orders in a manner that, in the market conditions of 2012, resulted in Knight buying stocks at the offer price — the price at which sellers were willing to sell — and selling them at the bid price — the price at which buyers were willing to buy. The offer price is, by definition, higher than the bid price. Buying at the offer and selling at the bid means buying high and selling low. It means losing money on every trade.

The function did this at machine speed. Thousands of orders per second, each one executed, each one losing a fraction of a cent per share, the fractions accumulating into dollars, the dollars accumulating into hundreds of dollars, the hundreds accumulating into thousands. The accumulation was instantaneous. In the first minute of trading, Knight's systems sent hundreds of thousands of orders and executed tens of thousands of trades. In the first five minutes, the volume had grown to millions of orders and hundreds of thousands of trades. The positions — the inventory of shares that Knight had bought and was now holding — were accumulating at a rate that the firm's risk management systems were not designed to process in real time.

Knight's trading desk, staffed by human beings who were monitoring the firm's positions on screens, noticed the accumulation within the first few minutes. The screens showed positions building in stocks across the market — positions that the desk had not intended to take, in sizes that the desk had not authorized. The desk's first response was to attempt to identify the source of the orders. The desk contacted Knight's technology team. The technology team began to investigate. The investigation, however, was conducted on a human timescale — minutes, not microseconds — and the system that was generating the losses was operating on a machine timescale. By the time the technology team had identified the problem, the losses had already become substantial.

The technology team's investigation was hampered by the complexity of Knight's systems and by the nature of the error. The error was not in the new SMARS code — the code that the team had deployed the previous evening. The error was in the old Power Peg code — the code that the team had not touched, that had been sitting on the server for nine years, and that the team had no reason to suspect. The team's investigation naturally focused on the new code, because the new code was what had changed. The investigation of the new code revealed no errors, because there were no errors in the new code. The error was in the interaction between the new code and the old code — the flag that the new code set, which reactivated the old code. The interaction was not something that the team's investigation was designed to detect, because the interaction had not been anticipated by anyone who had designed the deployment process.

The kill switch — the mechanism that would allow Knight's staff to halt trading on the malfunctioning system — was, in the architecture of Knight's risk management, a blunt instrument. The kill switch that existed was designed to halt all of Knight's trading, across all of its systems, in the event of a catastrophic failure. Pulling the kill switch would stop the losses, but it would also stop Knight's legitimate trading — the market-making activity that generated the firm's revenue, that served the firm's customers, and that the firm was obligated to provide to the exchanges. Pulling the kill switch was, in the language of risk management, a last resort. The staff's instinct was to find the specific system that was malfunctioning and to stop it, rather than to halt all of the firm's trading. The instinct was reasonable, but the execution was difficult, because the staff could not, in real time, identify which of the eight server groups was generating the rogue orders.

The minutes that followed were, by the accounts of those who were present, among the longest in the lives of the people who experienced them. The trading desk watched the positions build. The technology team searched for the source. The risk management team monitored the accumulating losses and the declining capital position. Tom Joyce, who had been in his office when the market opened, was informed within the first ten minutes. He came to the trading floor. The scene that greeted him was, by all accounts, controlled but tense — a room full of experienced professionals who understood that something was very wrong, who were working to identify what it was, and who were, in the meantime, watching the firm's capital disappear on the screens in front of them. Joyce's own response, as described by colleagues, was characteristic: he asked precise questions, he demanded specific information, and he did not — at least not visibly — panic. The composure was, in the circumstances, remarkable. It was also, in the end, insufficient, because the thing that was killing his firm was not susceptible to composure or to precise questions. The thing that was killing his firm was a piece of code, executing on a server, at a speed that no human intervention could match.

The search for the malfunctioning server group was the most frustrating aspect of the morning. Knight's systems generated logs — records of the orders they sent, the trades they executed, and the messages they exchanged with the exchanges. The logs were, however, voluminous, and they were distributed across the eight server groups. Searching the logs, in real time, for the source of the rogue orders, was a task that required both technical expertise and time. The technology team had the expertise. It did not have the time. The losses were accumulating at a rate of roughly ten million dollars per minute. Every minute that the team spent searching the logs was a minute in which the firm's capital declined by another ten million dollars. The team searched as fast as it could. The searching was fast. It was not fast enough.

At approximately ten-fifteen — forty-five minutes after the market had opened — Knight's technology team identified the eighth server group as the source of the rogue orders. The team manually removed the SMARS software from the eighth server group, which had the effect of stopping the Power Peg function. The removal was a simple operation — a technician accessed the server, deleted the software, and confirmed that the rogue orders had stopped. The simplicity of the fix was, in retrospect, the most painful aspect of the incident. The fix took, perhaps, thirty seconds. The search for the fix had taken forty-five minutes. In those forty-five minutes, the firm had lost $440 million — roughly ten million dollars per minute, or roughly one hundred and sixty-five thousand dollars per

second. The firm had been killed, and the killing had been done by a piece of code that could have been stopped, at any point during those forty-five minutes, by a thirty-second operation that no one had known to perform.

The losses were $440 million. Knight had lost, in forty-five minutes, roughly four times the amount it had earned in the entire previous year. The firm's capital, which had been sufficient to support its market-making activities at the start of the day, was now insufficient to cover the losses. Knight was, by any reasonable accounting, insolvent. The firm could not meet its obligations to its counterparties. It could not continue to make markets. It could not, in the ordinary course, open for business the next day.

The response, over the next several hours, was an exercise in crisis management. Knight's management contacted the SEC, the Financial Industry Regulatory Authority, and the various exchanges on which the firm operated, to inform them of the situation. The firm's clearing bank, Jefferies, was contacted to arrange emergency financing. The firm's counterparties — the other market makers, the exchanges, the clearinghouses — were informed, because each of them had exposure to Knight and needed to assess their own positions. The market, which had been operating normally throughout the forty-five minutes, was now aware that something was wrong. Knight's stock, which had opened the day at roughly eleven dollars, dropped by thirty-three percent over the course of the morning, as the news spread.

The rescue, when it came, was arranged over the weekend of August 4 and 5. The heads of Jefferies, TD Ameritrade, Getco, Blackstone, Stifel, and Stephens — a group of firms that included Knight's competitors, its customers, and its counterparties — gathered, in person and by phone, to decide whether to provide Knight with the emergency capital it needed to survive. The negotiations were, by all accounts, difficult. The firms were being asked to put up four hundred million dollars, in exchange for seventy percent of Knight's equity, at a price of one dollar and fifty cents per share — a price that represented an eighty-six percent discount to Knight's pre-crisis stock price. The firms were, in effect, being asked to rescue a competitor at a price that valued the competitor at a fraction of its recent worth.

The negotiations were complicated by the fact that the firms involved had, in many cases, conflicting interests. TD Ameritrade was one of Knight's largest customers — the retail brokerage routed a substantial fraction of its order flow to Knight for execution. If Knight failed, TD Ameritrade would need to find alternative arrangements for routing its orders, and the alternatives would, in the short term, be more expensive and less reliable. Getco was a direct competitor — a high-frequency trading firm that competed with Knight for market-making business. If Knight failed, Getco would inherit a portion of Knight's market share, but it would also inherit the disruption that the failure would cause in the market. Blackstone was a private equity firm with an interest in the broader stability of the financial system. Stifel and Stephens were regional financial services firms that had, over the years, competed with Knight in various business lines. Each firm had its own reasons to participate or to decline. Each firm's decision depended, in part, on what the other firms decided. The negotiation was, in the language of game theory, a coordination problem — one that was solved, over the course of the weekend, by the firms' collective recognition that the cost of letting Knight fail exceeded the cost of saving it.

The firms agreed. The agreement was announced on Monday, August 6. Knight survived, as a going concern, though its shareholders — who had owned a company worth roughly a billion dollars on July 31 — now owned a company worth roughly three hundred million dollars on August 6. Tom Joyce, who had run the firm through the crisis with a composure that was widely admired in the industry, remained as chief executive for the transition. The firm's market-making operations continued, serving the customers who depended on them. The system, in the narrow sense, held.

The firms that participated in the rescue did not do so out of charity. They did so because the alternative — letting Knight fail — would have imposed losses on each of them. Knight, as a market maker, had outstanding positions with every major participant in the U.S. equity market. If Knight had failed, those positions would have had to be unwound, in a market that was already stressed, and the unwinding would have imposed losses on the counterparties. The losses would not, in aggregate, have been as large as the losses that Bear Stearns or Lehman Brothers had imposed in 2008 — Knight was a much smaller firm — but they would have been substantial, and they would have fallen on firms that were, themselves, operating on thin margins. The rescue was, in this sense, an act of collective self-preservation. The firms saved Knight because the cost of saving Knight was less than the cost of letting Knight die.

The SEC, in the aftermath of the incident, conducted an investigation and, in October 2013, settled with Knight for twelve million dollars. The settlement was for failure to maintain adequate risk controls — a charge that Knight, which had already lost $440 million and 70% of its equity, was in no position to contest. The twelve million dollar fine was, in the context of the losses, a formality. The real punishment had been self-administered, by the firm's own systems, in the forty-five minutes before anyone could stop them.

The SEC's investigation, the findings of which were published in the settlement order, provided the first detailed public account of what had happened on the morning of August 1. The order described the deployment process, the Power Peg code, the flag that reactivated it, the forty-five minutes of uncontrolled trading, and the eventual identification and removal of the malfunctioning software. The order found that Knight had failed to maintain adequate risk controls — that the firm's pre-trade limits, post-trade monitors, and kill switches had been insufficient to prevent or to contain the losses. The finding was, technically, correct. The firm's risk controls had, in fact, been insufficient. As a broader matter, however, the finding was a statement of the obvious — an observation, made after the fact, that the controls had not worked, accompanied by a fine that was, in the context of the losses, a rounding error. The SEC's order did not, and could not, address the underlying problem, which was that the speed of Knight's trading systems had exceeded the speed of its risk controls, and that no set of controls that operated on a human timescale could have stopped the losses in time.

The SEC also, in the years that followed, promulgated a new regulation — Regulation SCI, for Systems Compliance and Integrity — that imposed stricter requirements on the technology and risk management practices of exchanges, clearinghouses, and large market makers. The regulation was, in part, a response to the Knight incident, and to a series of other technology-related market disruptions that had occurred in the preceding years, including the Flash Crash of May 2010, in which the Dow Jones Industrial Average had dropped nearly a thousand points in a matter of minutes before

recovering. The regulation required the firms it covered to maintain comprehensive testing procedures, to have effective kill switches, and to report significant disruptions to the SEC. The regulation was sensible. Whether it would prevent the next technology-related market disruption was, and remains, an open question. The regulation addressed the specific failures that the Knight incident had exposed — the inadequate testing, the absent kill switch, the insufficient risk controls. It did not, and could not, address the underlying dynamic, which was that the speed of the systems had permanently exceeded the speed of the controls, and that the gap between the two was widening.

The Knight Capital incident is the purest example so far of the velocity mismatch that the previous chapters have begun to trace. In the Silicon Valley Bank case, the depositors moved faster than the FDIC's resolution process. In the M-Pesa case, the customers adopted faster than the Central Bank of Kenya's regulatory process. In the Tether case, the issuance of dollar-denominated liabilities grew faster than any regulator's ability to supervise the issuer. In the Wirecard case, the distribution of accurate information moved faster than the regulator's ability to process it. In the Knight Capital case, the trading system moved faster than the risk management system. The firm's algorithms executed trades in microseconds. The firm's risk managers operated on a human timescale. The gap between the two timescales was the gap in which the $440 million was lost.

Knight's risk controls were not absent. The firm had pre-trade limits, post-trade monitors, and kill switches. The risk controls were reasonable. The problem was that the risk controls operated on a timescale that the trading systems had outgrown. The trading systems could lose $440 million in forty-five minutes. The risk controls could detect the losses and alert the risk managers in — well, in something less than forty-five minutes, but not in time to prevent the losses from accumulating to the point of insolvency. The risk controls were fast. The trading systems were faster. The gap between "fast by human standards" and "faster" was the gap in which Knight died.

The institutions that operate the world's financial infrastructure face a problem that better controls cannot solve. The speed of the systems has permanently exceeded the speed of the controls. Controls can be improved — kill switches made more granular, pre-trade limits more sophisticated, post-trade monitors more sensitive — but the controls will always be, by their nature, reactive. They detect conditions that the trading systems have already produced, and they intervene after the conditions have arisen. The trading systems, by contrast, are proactive — they produce the conditions, at a speed the controls cannot match. The asymmetry is structural, and it cannot be eliminated by better controls, because the controls will always be one step behind the systems they are designed to control.

The market, in the years since the Knight Capital incident, has continued to accelerate. The speeds at which trading systems operate have increased, the complexity of the systems has grown, and the number of venues, order types, and participants has expanded. The risk controls have, in many cases, improved — the Reg SCI requirements have forced firms to invest in their technology and their risk management, and the industry's own experience with incidents like Knight's has produced a heightened awareness of the risks. The improvement, however, has been incremental. The acceleration has been exponential. The gap between the speed of the systems and the speed of the controls has, if anything, widened. The next Knight Capital incident will, when it comes, involve a larger loss, sustained in a shorter period, because the systems will be faster and the controls will be, relatively, slower.

Knight Capital, in the end, did not die. It was rescued, at a price, by the firms that depended on it. The firm was subsequently acquired by Getco, in a merger that created KCG Holdings, which was itself later acquired by Virtu Financial. The Knight name, which had been one of the most respected in the market-making business, disappeared from the industry through a sequence of corporate combinations. Tom Joyce, who had run the firm through the crisis, departed. The employees who had been there on the morning of August 1, 2012, scattered to other firms. The eighth server group, the one that had run the Power Peg code, was decommissioned. The Power Peg code itself was, presumably, finally removed from Knight's codebase — the landmine, at last, defused.

The incident left, however, a mark on the industry that has not faded. The image of a firm losing $440 million in forty-five minutes, without a single human being making a conscious decision to take a risk, has become a touchstone for the discussion of technology risk in financial markets. The image is invoked whenever a trading firm experiences a technology glitch, whenever an exchange suffers an outage, whenever a flash crash raises questions about the stability of the market's structure. The image is invoked because it captures, in a single event, the essential vulnerability of a market that runs on software — the vulnerability that the software, which is written by human beings and tested by human beings and supervised by human beings, can do things that the human beings did not anticipate, at speeds the human beings cannot match, before the human beings can stop it.

The market runs on software now, and the software runs faster than the people who are supposed to be managing it. The people, on the morning of August 1, 2012, did what people do — they investigated, they deliberated, they decided, they acted. The software, in the same forty-five minutes, did what software does — it executed, relentlessly, the instructions it had been given, without pause, without hesitation, without the capacity to notice that the instructions were wrong. The instructions were wrong because of a flag in a directory on a server in Jersey City. The flag had been set by a deployment script that was supposed to have removed it, the deployment script had been run by a technology team that had every reason to believe the deployment was routine, and the routine had been performed correctly according to the procedures that governed it. The procedures were wrong, and they had been written, years earlier, by people who had not anticipated that a retired piece of code would be reactivated by a new deployment. The people who wrote the procedures were competent. The competence was not enough.

The institutions that operate the world's financial infrastructure — the banks, the regulators, the exchanges, the market makers, the auditors — are staffed by competent people. The people do their jobs, according to the procedures that govern their jobs. The procedures are written by competent people, who anticipate the risks they can imagine. The risks they cannot imagine — the risks that arise from the interaction of systems, from the speed of execution, from the accumulation of small errors into catastrophic losses — are the risks that the procedures do not cover, and the risks that the procedures do not cover are the risks that materialize. When they materialize, the institutions respond, after the fact, with new procedures designed to prevent the specific risk that just occurred. The new procedures will prevent that specific risk. They will not prevent the next one, which will be different, which will arise from an interaction that no one anticipated, at a speed that no one can match, in a system that no one fully understands.

Knight Capital lost $440 million in forty-five minutes because a piece of code that had been retired in 2003 was reactivated by a deployment in 2012. The code did what it was designed to do, the deployment did what it was designed to do, and the risk controls did what they were designed to do. Everything worked as designed. The design was wrong, because it did not contemplate the interaction between the old code and the new deployment. The interaction was not contemplated because no one had thought to contemplate it — and no one had thought to contemplate it because the interaction was obscure. It was the kind of thing that only becomes visible after it has happened, when the losses have been counted and the post-mortem has been conducted. The post-mortem was conducted. The losses were counted. The procedures were updated. The next interaction, the one that has not yet been contemplated, is waiting, somewhere, in a directory on a server, for the deployment that will reactivate it.

Chapter 6

The Neutral Utility That Wasn't

On the afternoon of February 26, 2022, two days after Russian tanks had crossed the border into Ukraine, the leaders of the United States, the European Commission, France, Germany, Italy, the United Kingdom, and Canada issued a joint statement from their respective capitals. The statement announced, among other measures, that the signatory governments had decided to disconnect "selected Russian banks" from SWIFT. The statement was brief, as joint statements of this kind tend to be. It did not specify which Russian banks would be disconnected. It did not specify when the disconnection would take effect. It did not specify what the practical consequences would be for the Russian economy, or for the global financial system, or for the banks' customers, or for the energy trade between Russia and Europe. It said, in substance, that the signatories had decided to use SWIFT as a weapon, and that the details would follow.

SWIFT — the Society for Worldwide Interbank Financial Telecommunication — is, despite its name, a relatively small organization. It is headquartered in La Hulpe, a suburb of Brussels, in a low-rise office complex that resembles, by the accounts of journalists who have visited it, a slightly downmarket community college or a regional insurance office. The building is set back from the road, surrounded by parking lots and landscaped berms, and is marked, at the entrance, by a modest sign bearing the SWIFT logo. The building is not, by any reasonable assessment, the kind of building from which one would expect the financial weapon of the century to be administered. It is, in all visible measures, the kind of building that houses mid-level bureaucrats, regional sales teams, and the administrative staff of organizations whose work is important but not dramatic. The disjunction between the building's appearance and its function is the central fact of SWIFT's existence, and it is the disjunction that the events of February 2022 brought, for the first time, into full public view.

SWIFT was founded in 1973, by a consortium of 239 banks from 15 countries, as a cooperative that would provide a standardized electronic messaging service for cross-border bank payments. The problem that SWIFT was created to solve was, in the early 1970s, becoming acute. Cross-border payments, at the time, were conducted by means of telex messages — human-readable text messages sent over the telegraphic network, which were slow, error-prone, and insecure. A bank that wished to send money to a bank in another country would compose a telex message instructing the receiving bank to credit the beneficiary's account, would send the message over the telex network, and would then reconcile the transaction through correspondent banking accounts that the two banks maintained with each other. The system worked, in the sense that payments were eventually made. It did not work well. Telex messages were frequently garbled, were sometimes intercepted, and were always slow. The reconciliation process, which depended on the two banks' maintaining matching records of their correspondent accounts, was labor-intensive and prone to dispute.

SWIFT's founders — a group of large banks, principally European and American, that handled the bulk of the world's cross-border payments — proposed to solve this problem by building a dedicated,

standardized, secure messaging network that would replace the telex. The network would use a common message format — a structured, machine-readable text that specified the sender, the receiver, the amount, the currency, the beneficiary, and the purpose of the payment. The network would be operated as a cooperative, owned by its member banks, governed by a board elected from the membership, and headquartered in a neutral country — Belgium — that was, in the geopolitical context of the Cold War, acceptable to both the American and the European banks and, importantly, to neither the Soviet Union nor the Eastern bloc. The choice of Brussels was, in 1973, a choice that reflected the politics of détente, the economics of European integration, and the practical necessity of finding a jurisdiction that could host a financial utility without becoming entangled in the superpower rivalry.

The cooperative began live operations in 1977, with the first SWIFT messages sent between member banks. The network grew rapidly. By the 1990s, SWIFT had become the global standard for cross-border payment messaging, handling millions of messages per day, connecting banks in virtually every country, and providing the infrastructure on which the international payments system operated. By 2022, SWIFT connected more than eleven thousand financial institutions in over two hundred countries and territories, and handled roughly forty-five million messages per day — a volume that, in aggregate, represented a substantial fraction of the world's cross-border financial traffic. SWIFT was, by any reasonable measure, the most important piece of financial infrastructure that most people had never heard of.

The thing that SWIFT did — the thing that its eleven thousand member institutions used it for — was not, in any technical sense, the transfer of money. SWIFT transferred messages. A bank that wished to send money to a bank in another country would compose a SWIFT message — a formatted text, structured according to a standardized message type — and would send the message over the SWIFT network to the receiving bank. The message would instruct the receiving bank to expect a payment, would specify the amount and the currency, would identify the beneficiary, and would provide the reference information necessary for the receiving bank to credit the beneficiary's account. The money itself — the actual settlement — would happen through correspondent banking accounts that the two banks maintained with each other, or through intermediary banks that provided correspondent services. SWIFT was, in essence, a very reliable, very expensive group chat for banks. The money moved through separate channels. SWIFT moved the instructions.

This distinction — between the messaging layer and the settlement layer — is the distinction that the Western governments, in their February 2022 statement, were exploiting. Cutting a bank from SWIFT did not prevent the bank from sending money abroad. The bank could still send money through correspondent banking relationships, through third-country intermediaries, through cash couriers, through hawala networks, or through cryptocurrency. What the cut prevented was the efficient, standardized, reliable communication of payment instructions. Without SWIFT, a Russian bank that wished to send money to a bank in Germany could not send a formatted SWIFT message instructing the German bank to expect the payment. It would have to communicate the instruction by other means — by fax, by email, by telephone, by a message sent over SPFS (the Russian domestic system) to a third-country bank that would then relay the instruction to the German bank through SWIFT. The payment could still be made. It would be slower, more expensive, more opaque, and more uncertain.

The cut from SWIFT did not stop the money. It made the money harder to move.

SWIFT, for the forty-nine years of its existence prior to February 2022, had maintained a posture of strict political neutrality. The cooperative's public statements, its corporate communications, and its leadership's public appearances all emphasized the same message: SWIFT was a utility, like the postal service or the telephone network, and it should not be used as an instrument of foreign policy. The message was accepted without serious challenge, because there was no serious challenge to accept. The Cold War did not produce a SWIFT exclusion, because the Soviet bloc maintained its own payments system. The post-Cold-War era did not produce a SWIFT exclusion, because the major powers, in the era of globalization, had no interest in disrupting the cross-border payments system on which their own economies depended. SWIFT's neutrality was a fiction that cost nothing to maintain, because it was never tested.

The first test came in 2012, when the European Union, acting under sanctions related to Iran's nuclear program, directed SWIFT to disconnect designated Iranian banks from its network. SWIFT complied. The disconnection was the first time in SWIFT's history that the cooperative had been used as a sanctions instrument, and it established a precedent that the cooperative's leadership had, for decades, insisted would never be established.

The Iranian disconnection was a manageable precedent. Iran was a relatively small economy, relative to the global financial system. The disconnection affected a limited number of banks, and the geopolitical consensus behind the sanctions was broad — the United States, the European Union, and most of the international community were aligned in their opposition to Iran's nuclear program. The disconnection was, in its practical effects, substantial for Iran — Iranian banks found it more difficult to send and receive international payments, and the cost of the alternatives was high — but it was minimal in its effects on the broader financial system. SWIFT continued to operate. The cooperative's other members continued to use the network. The precedent was set, but the precedent was, at the time, regarded as a special case — a response to a specific threat, applied to a specific country, under a specific set of international sanctions. The precedent was not regarded, in 2012, as the beginning of a new era in which SWIFT would be routinely used as a foreign policy tool.

The precedent was, however, more significant than it appeared. Before the Iranian disconnection, SWIFT had maintained its neutrality through a simple argument: the cooperative was a utility, subject to the law, and it would comply with whatever legal obligations were imposed on it. The argument had been tested, in the 2000s, by the controversy over the U.S. Treasury's Terrorist Finance Tracking Program, under which the Treasury had issued subpoenas for SWIFT transaction data as part of its counterterrorism efforts. The program, which had been conducted in secret since 2001, was exposed in 2006 by the New York Times, and produced a transatlantic controversy over the extent to which a Belgian cooperative should be providing transaction data to the U.S. government. The controversy was resolved, in 2010, by an EU-U.S. agreement that formalized the conditions under which SWIFT would provide data to the Treasury. The resolution was a model of the cooperative's approach to political pressure: SWIFT did not resist the pressure, but it negotiated the terms under which it would comply, and it obtained a legal framework that defined the scope of its cooperation. The approach preserved the fiction of neutrality — SWIFT was complying with the law, not making political choices — while

acknowledging that the cooperative was subject to the political direction of its most powerful members.

The Iranian disconnection stretched the fiction further. The disconnection was not, legally, a choice — SWIFT was subject to EU law, and the EU had designated the Iranian banks. But the disconnection was, politically, a choice — the EU could have chosen other sanctions, or could have chosen not to sanction Iran at all. The disconnection was a political act, implemented through a legal mechanism, by a cooperative that claimed to be neutral. The fiction held, but it was thinner than it had been. The precedent was set. SWIFT had been used as a weapon. The fiction of neutrality had been breached. The breach, in 2012, was small. In 2022, it would be large.

* * *

To understand why the SWIFT exclusion mattered, and why it mattered more than the exclusion of, say, a country from the international telephone network, it helps to understand what SWIFT actually does in the daily operation of the global financial system.

Cross-border payments, in the era before SWIFT, were conducted through a system of correspondent banking. A bank in one country that wished to send money to a bank in another country needed to have a relationship with the receiving bank — an account, maintained at the receiving bank, in which the sending bank held deposits in the receiving bank's currency. When the sending bank's customer wished to make a payment to the receiving bank's customer, the sending bank would instruct the receiving bank to debit the sending bank's correspondent account and to credit the beneficiary's account. The instruction was communicated by telex. The money was settled by adjusting the balances in the correspondent accounts.

The system worked, but it was cumbersome. Each bank needed to maintain correspondent relationships with banks in every country to which its customers might wish to send money. A large international bank might maintain hundreds of correspondent relationships, each requiring the maintenance of accounts, the reconciliation of balances, and the management of credit risk. Smaller banks, which could not maintain hundreds of correspondent relationships, relied on larger banks to act as intermediaries — routing their payments through the correspondent networks of the major international banks, which charged fees for the service. The system was a hierarchy, with the major international banks at the top and the smaller banks at the bottom, and with the costs of cross-border payments flowing downward, to the customers who ultimately paid the fees.

SWIFT did not replace this system. SWIFT replaced the messaging layer — the telex — that the system used to communicate payment instructions. The correspondent banking relationships, the accounts, the reconciliation, and the settlement all remained. What changed was the speed, reliability, and standardization of the messages. A SWIFT message, unlike a telex, was structured — it conformed to a defined format that specified, in a defined order, the sender, the receiver, the amount, the currency, the beneficiary, the correspondent banks involved, and the purpose of the payment. The format was

machine-readable, which meant that the receiving bank's systems could process the message automatically, without human intervention. The message was transmitted over a dedicated, secure network, which meant that it could not be intercepted or altered in transit. The message was confirmed, which meant that the sending bank knew, within seconds, that the receiving bank had received it.

The standardization was, in retrospect, the most important thing SWIFT provided. Before SWIFT, each bank had its own telex format, its own conventions for specifying amounts and beneficiaries, and its own procedures for reconciling messages with payments. A payment from a bank in New York to a bank in Frankfurt might arrive with the beneficiary's name spelled in a way the Frankfurt bank did not recognize, or with the amount specified in a format the Frankfurt bank's systems could not parse, or with the correspondent account number missing or incorrect. The payment would then require manual intervention — a human being at the Frankfurt bank would have to read the telex, figure out what it meant, and post the transaction manually. The manual intervention was slow, expensive, and error-prone. It was, in the aggregate, one of the principal costs of cross-border banking.

SWIFT solved this by defining, through a process of industry consultation, a set of standardized message types — the MT series — that specified, for each kind of payment, the exact format that the message would take. An MT103 message, for example, was a customer credit transfer — a payment from one bank's customer to another bank's customer, in a specific currency, for a specific amount, with specific fields for the beneficiary, the remitter, the correspondent banks, and the purpose. The format was the same, whether the payment was from New York to Frankfurt, from Tokyo to São Paulo, or from London to Mumbai. A bank that received an MT103 knew, without reading the message, what each field meant and how to process it. The standardization allowed banks to automate their payment processing, which reduced costs, reduced errors, and reduced the time it took for cross-border payments to settle. The standardization was, in the literal sense, the thing that made modern cross-border banking possible.

The network effects of SWIFT were, from the beginning, powerful. Each new bank that joined the network made the network more valuable to the banks already on it, because each new bank was a potential counterparty for the existing members. The network effects were also self-reinforcing: as more banks joined SWIFT, the cost of not being on SWIFT increased, because the banks that were on SWIFT could process payments more quickly and cheaply than the banks that were not. By the 1990s, SWIFT had achieved near-universal adoption among the world's banks. A bank that was not on SWIFT was, for practical purposes, cut off from the international payments system. It could send and receive payments through correspondent banks, but the process was slow, expensive, and unreliable — the same process that SWIFT had been built to replace. The choice, for any bank that wished to participate in cross-border finance, was not whether to join SWIFT. The choice was whether to participate in cross-border finance at all.

The governance of SWIFT reflected its cooperative structure. The cooperative was owned by its member banks, which elected a board of directors from among their number. The board set policy, approved the budget, and appointed the chief executive. The cooperative was incorporated under Belgian law, and was subject to the regulatory authority of the Belgian government, which had jurisdiction over its operations. The governance was, by design, decentralized and democratic — each

member bank had a stake in the cooperative, and the cooperative's decisions were made, in principle, by the membership. In practice, the large international banks — the American, European, and Japanese banks that handled the bulk of the world's cross-border payments — exercised disproportionate influence, because they provided the bulk of the cooperative's revenue and because their departure from the cooperative would have been catastrophic. The governance was, in this sense, a hierarchy dressed up as a cooperative — a structure that allowed the major banks to set policy while maintaining the fiction that the cooperative was a neutral utility serving all its members equally.

The neutrality of the cooperative was, for most of its history, maintained through a simple mechanism: SWIFT did what its member banks, and the governments that regulated them, told it to do. The cooperative did not set foreign policy. It did not decide which banks should be connected and which should not. It connected the banks that its members wanted connected, and it disconnected the banks that its members, or their governments, wanted disconnected. The neutrality was, in this sense, a neutrality of procedure rather than of substance. SWIFT would connect any bank that met its technical and membership criteria, regardless of the bank's nationality, ownership, or political alignment. It would disconnect any bank that the relevant authorities directed it to disconnect, regardless of the bank's nationality, ownership, or political alignment. The neutrality was in the process, not in the outcome. The process, for forty-nine years, produced outcomes that were consistent with the fiction. In 2012, the process produced an outcome that was not.

The 2012 Iranian disconnection established the precedent that SWIFT could be used as a sanctions instrument, but it did not, by itself, destroy the fiction of neutrality. The Iranian banks that were disconnected were designated under EU sanctions, and SWIFT, as an entity subject to EU law, was obligated to comply. The compliance was, legally, not a choice — SWIFT could not, under EU law, continue to provide services to designated entities. The disconnection was, in this framing, a legal obligation rather than a political decision. The framing preserved the fiction, barely, by allowing SWIFT to maintain that it was not making political choices — it was complying with the law.

The framing was tested again in 2014, when Russia annexed Crimea and the Western allies began discussing the possibility of a SWIFT exclusion for Russian banks. The discussion was, at the time, not pursued to implementation — the European governments, particularly Germany and Italy, were reluctant to disrupt the energy trade with Russia, and the SWIFT exclusion was regarded as a "nuclear option" that would be held in reserve. The Russian government, however, took the discussion seriously. The Central Bank of Russia, in the years after 2014, began developing the System for Transfer of Financial Messages — SPFS — as a domestic alternative to SWIFT. The development was, in the Russian official discourse, presented as a matter of national security: Russia could not, the argument went, remain dependent on a financial utility that was subject to the political direction of Western governments. The argument was, in retrospect, correct. The development of SPFS was, in retrospect, insufficient.

* * *

The implementation of the SWIFT exclusion began on March 2, 2022, six days after the invasion and four days after the joint statement. The European Union, acting through Council Regulation (EU) 2022/345, directed SWIFT to disconnect seven designated Russian banks from its network within ten days. The banks included VTB, Bank Otkritie, Sovcombank, Novikom, and several others. Sberbank, Russia's largest bank, was not initially included in the March 2 cutoff, but was added in June 2022. Gazprombank, the principal bank for Russia's gas exports, was not included in the cutoff at all — a deliberate exemption, made at the insistence of European governments that needed to continue paying for Russian gas through the banking system.

The exemption of Gazprombank was the most revealing moment in the sanctions story. The Western governments had decided to cut Russia off from the international payments system — to weaponize SWIFT, to breach the fiction of neutrality, to use the cooperative as an instrument of foreign policy. They had done all of this in the name of punishing Russia for its invasion of Ukraine. And then, in the same breath, they had exempted the bank through which Europe paid for Russian gas, because Europe needed the gas. The exemption said, more clearly than any official statement, that the SWIFT exclusion was not an absolute measure. It was a measure that would be applied where it did not impose unacceptable costs on the countries applying it. The weapon had a safety catch, and the safety catch was the energy trade.

SWIFT, for its part, implemented the exclusion as directed. The cooperative issued a brief statement confirming that it would comply with the EU regulation and would disconnect the designated banks within the specified timeframe. The statement was terse. It did not comment on the policy. It did not express an opinion on the weaponization of the cooperative's network. It did not lament the breach of the neutrality fiction. It said, in substance, that SWIFT was a law-abiding entity subject to EU law, and that it would do what the law required. The terseness was an acknowledgment that the cooperative had no agency in the matter. SWIFT was not consulted. SWIFT was not asked for its opinion. SWIFT was told.

The practical consequences of the exclusion were immediate and substantial for the Russian banks and the Russian economy. The designated banks could no longer send or receive SWIFT messages. Their correspondent banks, in Europe and elsewhere, could no longer receive payment instructions through the standard channel. Payments that had previously taken hours to settle now took days, or weeks, or did not settle at all. Russian importers, who depended on SWIFT to pay foreign suppliers, found that their payments were being returned, delayed, or refused by correspondent banks that were unwilling to process Russian transactions even through non-SWIFT channels. Russian exporters, who depended on SWIFT to receive payment for their goods, found that their foreign customers were unable to send payments through the standard channel, and were reluctant to use alternative channels that were slower, more expensive, and less certain.

The Russian government's response, in the immediate aftermath of the exclusion, was to accelerate the deployment of SPFS. The system, which had been operational since 2017 but which had been used primarily for domestic traffic, was expanded to handle a larger volume of messages, and Russian banks

were directed to route their domestic payments through SPFS rather than SWIFT.

The SPFS system deserves a moment of description, because it illustrates both the ambition and the limits of Russia's preparation. SPFS — the System for Transfer of Financial Messages — was built by the Central Bank of Russia as a domestic equivalent of SWIFT. The system used a similar message structure, based on the same ISO standards that SWIFT used, and was designed to be interoperable, at the message level, with SWIFT's formats. A Russian bank that sent an SPFS message was, in effect, sending a message that was structurally identical to a SWIFT message, over a different network. The interoperability at the message level was intended to make it easier for Russian banks to switch from SWIFT to SPFS, and to make it easier for foreign banks to connect to SPFS if they wished to receive messages from Russian counterparties. The interoperability was, however, only at the message level. The network itself was separate, and the network's reach was limited to the banks that had chosen to connect to it.

By 2022, SPFS had roughly four hundred participating institutions. The majority were Russian banks, which had been directed by the Central Bank to connect. A smaller number were banks from Belarus and Kazakhstan, which had joined as part of the Eurasian Economic Union's payments integration. A handful of banks from other countries — largely from jurisdictions that maintained friendly relations with Russia — had also connected, but the number was small, and the volume of international traffic that flowed through SPFS was minimal. The system was a capable piece of infrastructure. It was also a regional network with limited international reach. The gap between the architecture and the adoption was the gap that the exclusion exposed.

The expansion was successful — SPFS handled the domestic traffic that had previously gone through SWIFT, and Russian domestic payments continued to function. A Russian company paying a Russian supplier, or a Russian individual transferring money between Russian bank accounts, experienced no disruption. The domestic payments system, which had been partially dependent on SWIFT for messaging, migrated to SPFS without significant incident. The Central Bank of Russia, which had been preparing for this contingency since 2014, executed the migration with a competence that surprised some Western observers. The preparation had not been wasted. The domestic system held.

The international traffic, however, was a different matter. SPFS was not interconnected with the international payments system. A Russian bank could send an SPFS message to another Russian bank, but it could not send an SPFS message to a bank in Germany, or China, or Turkey. The international payments required either a SWIFT connection (which the designated banks no longer had) or a bilateral arrangement with a correspondent bank in a third country.

The third-country correspondent banks — in Turkey, the United Arab Emirates, China, and a handful of other jurisdictions — became, in the months after the exclusion, the principal channel through which Russian international payments were routed. A Russian importer that wished to pay a supplier in Europe could route the payment through a bank in Turkey: the Russian bank would send an instruction (by email, by fax, or through SPFS if the Turkish bank was connected) to the Turkish bank, the Turkish bank would send a SWIFT message to the European bank, and the European bank would credit the supplier's account. The payment would settle, eventually, with the Turkish bank taking a fee

for the intermediation. The system was slower, more expensive, and more opaque than the direct SWIFT connection that had previously been used. It was, however, functional. The money moved. The trade continued.

The Turkish and Emirati banks, in the months after the exclusion, found themselves in a position that was, from their perspective, lucrative and uncomfortable. The fees they charged for intermediating Russian payments were substantial — higher, by several multiples, than the fees they charged for ordinary cross-border payments. The volume of Russian payments routed through their systems was also substantial, as Russian importers and exporters sought alternative channels for their international transactions. The profitability was attractive. The discomfort came from the Western governments, which made it clear that they regarded the intermediation as a potential sanctions evasion risk, and that they expected the Turkish and Emirati banks to apply enhanced due diligence to Russian-related transactions. The banks were, in effect, caught between the profitability of the Russian business and the risk of Western sanctions, and they managed the tension by applying the due diligence selectively — processing the payments that were clearly legitimate, declining the payments that were clearly problematic, and making case-by-case judgments on the payments that fell in between. The case-by-case judgments were, in their aggregate effect, the mechanism through which the Russian international payments system continued to function. The mechanism was imperfect. The mechanism was sufficient.

The Chinese alternative — CIPS, the Cross-Border Interbank Payment System — became, in the months and years after the exclusion, an increasingly important channel for Russia-Asia payments. CIPS, which had been launched in 2016 as China's equivalent of SWIFT, had grown steadily, and by 2022 connected roughly fourteen hundred financial institutions in over a hundred countries. Russian banks, cut from SWIFT, began connecting to CIPS, routing their Asian payments through the Chinese system. The routing was a partial solution — CIPS handled RMB-denominated payments, and the majority of Russia's trade with Asia was denominated in RMB or in other non-dollar currencies. The system was, however, not a full substitute for SWIFT. CIPS did not have the universal reach of SWIFT, and it did not handle the full range of currencies and message types that SWIFT supported. It was, for Russia, a lifeline — but a lifeline that covered only a portion of the country's international payment needs.

The pattern that emerged, in the months and years after the exclusion, was one of fragmentation. Russian international payments, which had previously been routed through SWIFT to counterparties around the world, were now routed through a patchwork of channels: SPFS for domestic traffic, CIPS for Asian traffic, third-country correspondent banks for traffic to Europe and the Americas, and, in some cases, cryptocurrency for payments that could not be routed through any banking channel. The patchwork was functional — the payments moved, the trade continued, the Russian economy did not collapse. The patchwork was also a vindication of the thesis that SWIFT was the messaging layer rather than the settlement layer. The money had found other channels. The messages had found other channels. The system had fragmented, but it had not broken.

The fragmentation came at a cost. The costs were borne, in the first instance, by the Russian banks and their customers, who paid higher fees for slower payments. The costs were borne, in the second

instance, by the European businesses that traded with Russia, who found that their Russian customers could no longer pay them through the standard channel, and that alternative channels were unreliable. The costs were borne, in the third instance, by the international payments system itself, which had, for forty-nine years, operated on a single, universal messaging standard, and which now operated on a patchwork of regional and bilateral standards that were not fully interoperable. The costs were, in the aggregate, substantial. They were not, however, catastrophic — not for Russia, and not for the international payments system. The system fragmented. The system continued.

The weaponization of SWIFT is the chapter in which the velocity mismatch runs in both directions. In the previous chapters, the mismatch has run in one direction: the money (or the information, or the trading system) moved faster than the institution could respond. In the SWIFT case, the mismatch ran in both directions at once. SWIFT, which had spent forty-nine years building a reputation for neutrality, lost the reputation in a single weekend. The loss was faster than the cooperative could process — the board, the management, and the staff were, by all accounts, unprepared for the speed with which the Western governments moved from discussion to announcement to implementation. The fiction of neutrality, which had been maintained for half a century, was retired in forty-eight hours. The institution could not adapt fast enough to preserve the fiction, because the geopolitical pressure was greater than the institution's capacity for resistance.

Russia, on the other side, had spent eight years building SPFS, and the eight years were not enough. The Central Bank of Russia had recognized, in 2014, that SWIFT was a vulnerability — that the cooperative's neutrality was a fiction that could be breached, and that Russia's dependence on SWIFT was a strategic risk. The Central Bank had responded by building a domestic alternative, and the alternative had, by 2022, reached a level of maturity that allowed it to handle domestic traffic. The alternative had not, however, reached a level of maturity that allowed it to replace SWIFT for international traffic. The eight years of preparation were insufficient, because the task of building an international payments network — a network that would connect Russian banks to counterparties in Europe, Asia, and the Americas — was a task that could not be completed in eight years, or in sixteen, by a single country acting alone. SWIFT had taken decades to achieve universal adoption, and it had achieved universal adoption because it was a cooperative, built by its member banks, for the benefit of its member banks. SPFS was a national system, built by one country, for the benefit of that country's banks. The difference was structural, and it could not be overcome by additional years of development.

The mismatch, in the SWIFT case, was a mismatch between the speed of geopolitical events and the speed of institutional preparation. The invasion of Ukraine was the product of years of Russian strategic calculation. The Western response — including the SWIFT exclusion — was the product of weeks of diplomatic negotiation. The implementation was the product of days of regulatory drafting. The sequence, from invasion to implementation, took six days. The sequence, from the first serious discussion of a SWIFT exclusion (in 2014) to the implementation (in 2022), took eight years. The eight years were not enough. The six days were more than enough. The mismatch between the eight years of preparation and the six days of implementation is the mismatch.

SWIFT, in the aftermath of the exclusion, has continued to operate. The cooperative has not, as of the time of writing, been formally reorganized or restructured. Its headquarters remains in La Hulpe. Its

staff continues to maintain the network, process the messages, and provide the services that its members require. The cooperative's public communications have returned, after the brief flurry of statements in February and March 2022, to the quieter posture that characterized its pre-exclusion communications. The cooperative has reverted, in its public posture, to the fiction that it is a neutral utility, subject to the law, doing what it is told. The fiction is, however, no longer credible. SWIFT was used as a weapon. The weapon was used at the direction of Western governments, against a designated adversary, in furtherance of a specific foreign policy objective. The cooperative that maintains the weapon cannot, having used it, plausibly maintain that it is neutral. The neutrality was a fiction that survived forty-nine years of untested existence. It did not survive six days of actual use.

The consequences of the breach extend beyond Russia. Every country that maintains a SWIFT connection — which is to say, every country whose banks participate in international finance — now knows that the connection can be severed, at the direction of the Western governments, in the event of a geopolitical confrontation. The knowledge has, in the years since the exclusion, produced a series of responses. China has accelerated the development of CIPS, expanding its reach and its message types. India has explored the development of alternative payment corridors, particularly for trade with Russia and Iran. The BRICS countries — Brazil, Russia, India, China, and South Africa — have discussed, in their summits, the possibility of building a shared payments infrastructure that would reduce their dependence on SWIFT.

The BRICS discussions have been slow. The five countries have different interests, different financial systems, and different relationships with the Western governments that control SWIFT. China, which has the most developed alternative in CIPS, has been cautious about opening the system to broader participation, preferring to maintain control over a system that it built and that it operates. India, which has its own domestic payment systems but no international equivalent, has been interested in alternatives that would reduce its dependence on the dollar but wary of antagonizing the United States. Brazil and South Africa, which are smaller participants in the international financial system, have been supportive in principle but limited in their capacity to contribute. Russia, which is the most directly affected, has been the most vocal advocate for an alternative, but has the least to offer in terms of financial infrastructure. The discussions have produced declarations, working groups, and feasibility studies. They have not, as of the time of writing, produced a concrete alternative.

The discussions have, however, established that the search for an alternative is now a serious undertaking, conducted by serious institutions, with serious resources. The search is no longer a theoretical exercise, conducted by think tanks and academic conferences. It is a policy priority, pursued by central banks and finance ministries, in countries that collectively represent a substantial fraction of the world's population and economic output. The search will take years, or decades, to produce a result. The result, when it comes, will not be a single network that replaces SWIFT. It will be a patchwork — of CIPS, of SPFS, of bilateral arrangements, of regional clearing systems, of whatever else the countries that need to move money outside the SWIFT network devise. The patchwork will be slower, more expensive, and less reliable than the single network it replaces. It will, however, be outside the reach of the Western governments that weaponized SWIFT. The patchwork will be a form of independence for the countries that use it. The independence will come at a cost. The countries that use

the patchwork will pay the cost, because the alternative — remaining dependent on a utility that can be weaponized — is, in the world that February 2022 brought into being, no longer acceptable.

The search for an alternative is the most consequential outcome of the SWIFT exclusion. The cooperative's principal asset — the thing that made it indispensable, the thing that made the exclusion a weapon rather than an inconvenience — was its universality. SWIFT connected every bank to every other bank, through a single, standardized network. The universality was the product of decades of network effects, of the kind that accumulate when a standard becomes the standard. Before February 2022, that universality was regarded as unchallengeable. After February 2022, it is regarded as challengeable. The challenge will take years, or decades, to materialize — if it materializes at all. The challenge has, however, begun.

The institutions that manage the world's financial infrastructure — in this case, a Belgian cooperative that sends text messages between banks — were designed for a world in which the geopolitical environment was stable, in which the major powers shared an interest in maintaining the infrastructure, and in which the infrastructure's neutrality was a fiction that cost nothing to maintain. That world existed, in its most fully developed form, from the end of the Cold War to the invasion of Ukraine — a period of roughly thirty years, during which SWIFT grew to universal adoption and during which the fiction of neutrality was never seriously tested. The world that has replaced it is one in which the geopolitical environment is contested, in which the major powers do not share an interest in maintaining the infrastructure, and in which the infrastructure's neutrality is a fiction that can no longer be maintained. The institutions have not adapted to the new world, and they may not be able to adapt, because the adaptation required — the construction of a genuinely neutral, genuinely universal payments infrastructure, acceptable to all the major powers — is an undertaking that no single institution, and no single country, can accomplish alone.

SWIFT, in La Hulpe, continues to send messages. The messages still flow, between the banks that remain connected, in the formats that have been standardized over fifty years. The network still works. The cooperative still maintains, in its public posture, that it is a neutral utility. The posture is, for the moment, sufficient, because the vast majority of the world's banks continue to use SWIFT, and because no adequate substitute yet exists. The posture will remain sufficient until a substitute emerges, or until the next geopolitical confrontation produces the next exclusion.

The headquarters in La Hulpe, in the meantime, continues to house the cooperative that maintains the network. The building has not changed. The staff has not changed, in any visible respect. The modest sign at the entrance still bears the SWIFT logo. The parking lots are still there, and the landscaped berms, and the low-rise office complex that resembles a community college. It looks, to a passing observer, exactly as it looked before February 2022. The function of the building has, however, changed. The headquarters of a neutral utility has become the headquarters of a weapon. The weapon is sheathed, for the moment, in the procedures and the governance and the corporate communications of a Belgian cooperative. It will, at the direction of the governments that control it, be unsheathed again. The direction will come, when it comes, at the speed of a geopolitical decision. The cooperative will comply, when the direction comes, at the speed of a bureaucratic process. The gap between the two speeds is the gap in which the next exclusion will occur.

The building in La Hulpe is, in the end, the image that this chapter leaves. A modest office complex in a Belgian suburb, housing the most important financial utility of the late twentieth century, now housing the most consequential financial weapon of the early twenty-first. The utility was neutral. The weapon is not. The transition from one to the other took six days, and it will not be reversed.

Chapter 7

Two Days, Two Hundred and Fifty Percent

On the morning of March 8, 2022, the price of nickel on the London Metal Exchange began, shortly after the electronic trading session opened, to rise at a rate that the market's participants had not previously seen and that the exchange's systems had not been designed to process. Nickel had closed the previous day — Monday, March 7 — at roughly forty-eight thousand dollars per tonne, which was itself a substantial increase from the roughly thirty thousand dollars per tonne at which nickel had been trading the week before. The increase over Monday had been driven by the same force that was driving the increase on Tuesday: the Russian invasion of Ukraine, which had begun on February 24, and which had raised concerns about the availability of Russian nickel, which accounted for roughly seven to ten percent of global supply. The concerns were, in their substance, reasonable. Russia is a major nickel producer, and the invasion had raised the prospect of sanctions, of supply disruptions, and of a general reduction in the availability of Russian metals on the global market.

The concerns, however, did not, by themselves, explain the magnitude of the price move. Nickel had risen from thirty thousand to forty-eight thousand dollars over the course of a week — a sixty percent increase, which was large but not unprecedented for a commodity market responding to a supply shock. What happened on the morning of March 8 was of a different order. The price crossed sixty thousand dollars. Then seventy thousand. Then eighty thousand. Then ninety thousand. Then, briefly, it crossed one hundred thousand dollars per tonne — a price that nickel had never, in the history of the market, reached. The move from forty-eight thousand to one hundred thousand dollars had taken hours. The move from thirty thousand to one hundred thousand had taken two trading days. In those two days, the price of nickel had risen by roughly two hundred and fifty percent.

The cause of the spike was not, or not only, the Russian supply concern. The cause was a short squeeze. A short squeeze is a market dynamic that occurs when a participant has sold short — has sold, in effect, something it does not own, on the expectation that the price will fall and that the participant will be able to buy it back at a lower price — and when the price, instead of falling, rises. The rising price forces the short seller to post margin — to deposit cash or other collateral with the clearinghouse, to cover the unrealized losses on the short position. If the short seller cannot post the margin, the clearinghouse will close out the position — will buy the asset on the open market, at whatever price is available, to replace the short seller's obligation. The buying, in turn, pushes the price higher, which increases the losses, which triggers more margin calls, which forces more buying. The dynamic is self-reinforcing, and it can produce extraordinary price moves in a short period, because the buying that drives the price higher is not driven by genuine demand but by the mechanical necessity of closing out the short position.

The short seller, in this case, was Tsingshan Holding Group — a Chinese company, founded in 1988 by a man named Xiang Guangda, that had grown over three decades into the world's largest producer of stainless steel. Xiang, who was sixty-four years old at the time of the crisis, was known in

the Chinese metals industry by the nickname "Big Boss" — a nickname that captured both the scale of his enterprise and the directness of his management style. Tsingshan produced stainless steel from a raw material called nickel pig iron, which is a low-grade nickel ore that is processed, in Tsingshan's facilities in Indonesia and China, into a feedstock for stainless steel production. The company was, by most measures, a remarkably successful industrial enterprise, with revenues exceeding fifty billion dollars and operations spanning the global metals supply chain.

Tsingshan's nickel short position was a hedge. The company produced nickel, in the form of nickel pig iron, and the price of nickel pig iron was correlated, to a degree, with the price of pure nickel traded on the LME. A company that produces nickel and is concerned that the price of nickel might fall can hedge that concern by selling nickel short on the futures market — by committing, through a futures contract, to deliver nickel at a specified price at a specified future date. If the price of nickel falls, the short position generates a profit that offsets the reduced revenue from the company's physical nickel sales. If the price of nickel rises, the short position generates a loss, but the loss is offset by the increased revenue from the physical sales. The hedge, in textbook form, reduces the company's exposure to nickel price fluctuations, and is a standard practice in the commodities industry.

Tsingshan's hedge, however, was not textbook. The position was enormous — estimated, at the time of the crisis, at between one hundred thousand and one hundred and fifty thousand tonnes of nickel. The LME's deliverable inventory — the nickel held in LME-approved warehouses, available for delivery against futures contracts — was roughly one hundred thousand tonnes. Tsingshan's short position was, in other words, larger than the entire deliverable supply of nickel on the exchange. The size of the position meant that Tsingshan could not, if the price rose and the position needed to be closed out, simply buy nickel on the LME to cover the short. There was not enough nickel, in the LME's warehouses, to cover it. The position would have to be closed out through cash settlement — through Tsingshan paying the difference between the contract price and the market price, in cash, to the counterparties who held the other side of the trade. The cash settlement, if the price rose sufficiently, would be catastrophic.

The size of the position had been, for months before the crisis, a source of concern among some participants in the nickel market. Tsingshan had been building the position since at least mid-2021, on the thesis that nickel prices would fall as new Indonesian supply came online. The thesis was correct — nickel prices did fall, eventually, as Indonesian supply increased. But the position was so large that it had become, in the language of the market, "cornerable" — vulnerable to a squeeze by counterparties who knew that Tsingshan was short, who knew that the position was too large to cover through physical delivery, and who could, by buying nickel and driving up the price, force Tsingshan into a margin call that the company could not meet. The squeeze, when it came, was driven by a combination of genuine supply concern (the Russian invasion) and speculative buying by counterparties who recognized the vulnerability and who front-run Tsingshan's inevitable covering by buying ahead of it.

The counterparties who bought nickel on the morning of March 8, driving the price from forty-eight thousand toward one hundred thousand dollars, were a mix of hedge funds, proprietary trading firms, and other participants in the metals market. Some were buying because they believed the Russian supply disruption justified higher prices. Some were buying because they knew Tsingshan was

short and they wanted to be on the other side of the trade. Some were buying because they saw the price rising and did not want to miss the move. The motivations were mixed, but the effect was unified: the price rose, the margin calls mounted, and Tsingshan, sitting in its offices in Wenzhou, faced losses that, at the peak, would have exceeded the company's ability to pay.

Xiang Guangda, in Wenzhou, was reported by subsequent accounts to have been initially reluctant to reduce the position. The reluctance was, by the accounts of those who knew him, characteristic. Xiang had built Tsingshan over three decades by making large bets on the direction of metals prices — bets that had, more often than not, been correct, and that had made him one of the wealthiest men in China. The nickel short was another such bet. Xiang believed that nickel prices would fall, and he was willing to hold the position through adverse price moves, on the expectation that the market would eventually come around to his view. The expectation was, as it happened, correct — nickel prices did fall, eventually, after the crisis had passed. But the expectation was also, in the short term, catastrophically wrong, because the price rose far enough, and fast enough, to trigger margin calls that Tsingshan could not meet before the price fell again. The bet was right. The timing was not. The distinction between a correct bet and a correctly timed bet is, in a market that moves at electronic speed, the distinction between a profit and a bankruptcy.

The margin calls that Tsingshan faced on the morning of March 8 were, by the estimates that subsequently emerged, in the billions of dollars per hour. The clearinghouse, LME Clear, was issuing the calls through Tsingshan's clearing members — the banks and financial institutions that stood between Tsingshan and the clearinghouse, and that were responsible, in the first instance, for meeting the clearinghouse's margin requirements. The clearing members, in turn, were demanding that Tsingshan post the collateral. Tsingshan, whose assets were largely in physical nickel and in production facilities rather than in liquid cash, was struggling to meet the demands. The company had cash, but not enough. The company had nickel, but the nickel was in Indonesia and China, not in LME-approved warehouses, and could not be delivered against the LME contracts. The company had production facilities, but those could not be sold quickly enough to raise the required cash. Tsingshan was solvent on its balance sheet. It was illiquid. The distinction between solvency and liquidity is, in a margin call, the distinction between survival and failure. Tsingshan was solvent but illiquid, and the illiquidity was, on the morning of March 8, sufficient to destroy the company if the price did not stop rising.

* * *

The London Metal Exchange is, among the world's major commodities exchanges, the one that most visibly retains the architecture of an earlier century. The exchange was founded in 1877, in a coffeehouse above a hat shop in Lombard Street, in the City of London, by a group of metal merchants who needed a place to trade copper and tin. The merchants had, before 1877, conducted their business informally — in coffeehouses, in taverns, in the back rooms of metal brokerage firms. The founding of

the exchange was an effort to bring order to the trade, by establishing standard contract sizes, standard delivery dates, and a standard venue for price discovery. The order was successful. The LME became, over the course of the twentieth century, the global benchmark for industrial metals pricing — the exchange whose prices, set daily, were used by producers and consumers around the world to price their physical contracts.

The architecture of the LME reflected the industrial age in which it was built. The exchange's principal mechanism for price discovery was the Ring — a circle of traders, seated on red leather benches, who shouted bids and offers at each other in a designated ring, with official prices set twice daily, at one o'clock and two o'clock in the afternoon, London time. The Ring trading was supplemented, in the later twentieth century, by an electronic trading system, and by a telephone market for over-the-counter trading. But the Ring remained, into the twenty-first century, the exchange's defining institution — the place where the official prices were set, and the symbol of the exchange's continuity with its 1877 founding. The LME was, in 2022, the last exchange in Europe that still used open-outcry ring trading for its official price-setting. The combination of a nineteenth-century trading floor and a twenty-first-century electronic market produced a certain cognitive dissonance, but it also produced a market that worked, in the sense that it generated prices that the global metals industry accepted as authoritative.

The Ring itself deserves description, because it is the physical embodiment of the world the LME was built for, and because the contrast between the Ring and the electronic market that ultimately produced the March 8 crisis is the contrast between two eras of price discovery. The Ring is a circular space, roughly twenty feet in diameter, located in the LME's offices at 10 Finsbury Square, in the City of London. Around the inside of the circle are two tiers of red leather benches. On these benches sit the representatives of the LME's Ring dealing members — the firms that are licensed to trade in the Ring, and that include the major metals trading desks of global banks and commodities firms. Each firm has a designated seat. At the appointed hour, the traders take their seats, and the trading begins. A trader who wishes to buy shouts a bid. A trader who wishes to sell shouts an offer. The traders shout at each other, across the ring, for a period of five minutes per metal. At the end of the five minutes, the official price is set — the last bid and offer at which trades were executed. The process is repeated, for each metal, twice daily. The prices set in the Ring are the LME's official prices, and they are the prices used by the global metals industry to settle physical contracts.

The Ring is absurdly slow. Five minutes per metal, two sessions per day, a handful of traders shouting at each other across a circle of red leather benches. An electronic matching engine can execute thousands of trades per second. The Ring executes, at most, a few dozen trades per five-minute session. The Ring is also remarkably resilient. The traders in the Ring can see each other. They can read each other's body language. They can sense, in a way that an electronic matching engine cannot, when a counterparty is under stress, when a position is being forced, when a market is becoming disorderly. The Ring provides a form of market surveillance that electronic systems have never fully replicated. The traders in the Ring can, and do, slow down when the market is moving too fast. They can refuse to trade. They can ask each other what is happening. The Ring is a rate-limiting mechanism — a mechanism that prevents the market from doing things faster than the people running it can understand.

The electronic market, which has no such mechanism, does not slow down. It does what its algorithms tell it to do, at the speed of its network connections, regardless of whether the prices it produces are sustainable.

The LME's history, from 1877 to the present, is a history of an institution that adapted, gradually, to the changing needs of the metals industry, while retaining the core of its original architecture. The exchange added metals — aluminum, zinc, lead, and others — as the industrial demand for those metals grew. The exchange added electronic trading, in the 1990s and 2000s, as the technology became available and as the participants demanded the ability to trade outside the Ring's fixed hours. The exchange established a clearinghouse, LME Clear, in 2014, to replace the bilateral clearing arrangements that had previously governed the settlement of LME trades. Each adaptation was a response to the changing market — an effort to bring the exchange's architecture into alignment with the needs of its participants. Each adaptation was, however, an incremental change to an institution whose core assumptions — that prices would move gradually, that positions would be sized relative to deliverable supply, that the participants would be known to each other — remained unchanged. The adaptations kept the exchange current. They did not prepare it for a market in which a single participant could hold a position larger than the deliverable supply, and in which the price could move two hundred and fifty percent in two days.

The exchange's clearing and settlement infrastructure, like its trading floor, reflected its history. The LME used a clearinghouse — LME Clear, which was established as a wholly-owned subsidiary of the exchange in 2014 — to guarantee the performance of trades executed on the exchange. The clearinghouse stood between the buyer and the seller, assuming the counterparty risk: if the seller defaulted, the clearinghouse would pay the buyer; if the buyer defaulted, the clearinghouse would pay the seller. The clearinghouse managed this risk by requiring participants to post margin — cash or other collateral, deposited with the clearinghouse, in amounts that reflected the size of the participants' positions and the volatility of the market.

The margining system was designed for a market in which prices moved gradually. The LME's historical experience, over a century and a half of trading, was that industrial metals prices moved in increments of a few percent per day, with occasional larger moves of five or ten percent in response to significant supply disruptions. The margining system was calibrated to this experience. Participants were required to post initial margin — a deposit, set by the clearinghouse, that reflected the expected one-day price move — and variation margin — additional deposits, called intraday if the price moved against the participant, to cover the unrealized losses on the position. The variation margin calls were issued by the clearinghouse, received by the participant, and paid, in a process that typically took hours rather than minutes. The process was designed to work when prices moved gradually, because the gradual movement gave the clearinghouse time to issue the calls, gave the participants time to pay them, and gave the market time to absorb the positions of any participant that could not pay and that had to be closed out.

The system had worked, for the LME's first century and a half, without a major clearinghouse failure. It had weathered the oil shocks of the 1970s, the Asian financial crisis of the 1990s, the global financial crisis of 2008, and the European debt crisis of 2010. In each case, prices had moved, margin

calls had been issued, and participants had met the calls or had been closed out by the clearinghouse. The system had proven adequate to the risks that the industrial metals market had historically presented.

The system was not, however, designed for a market in which a single participant held a position larger than the entire deliverable supply, and in which the price could move two hundred and fifty percent in two days. The margining system's assumption — that prices would move gradually enough for margin calls to be issued, received, and paid before the next threshold was breached — was an assumption that held for normal market conditions. It did not hold for the conditions that obtained on the morning of March 8, 2022. On that morning, the price of nickel was moving not by percent per day but by percent per hour. The margin calls that the clearinghouse was issuing to Tsingshan were, by the time they were received, already insufficient — the price had moved further, and the required margin had increased, before the previous call could be paid. The clearinghouse was, in effect, chasing a moving target, and the target was moving faster than the clearinghouse's processes could follow.

The ownership of the LME, in 2022, added a further layer of complexity to the crisis. The exchange had been acquired in 2012 by Hong Kong Exchanges and Clearing — HKEX — for £1.39 billion. The acquisition was controversial. The LME was a small exchange, with a specialized membership and a distinctive culture. The acquisition by HKEX was driven by the belief that the LME's benchmark prices for industrial metals would become increasingly important as Asian — and particularly Chinese — demand for metals grew, and that HKEX, as the owner of the LME, would be positioned to benefit from that growth. The acquisition had, in the years since 2012, produced mixed results. The LME had continued to operate, under HKEX ownership, with a substantial degree of autonomy, and the synergies that HKEX had hoped for had been slow to materialize. The crisis of March 2022 would test the relationship between the LME and its Hong Kong parent, and would test, more broadly, the question of whether an exchange founded in 1877 to trade metals in a London coffeehouse could operate, in the twenty-first century, under Asian ownership, in a global market that moved at electronic speed.

The LME's membership, at the time of the crisis, was a mix of the world's largest metals trading firms, banks, and industrial participants. The exchange's clearing members — the firms that had direct clearing relationships with LME Clear, and that were responsible for meeting the clearinghouse's margin calls — included the major metals trading desks of global banks and commodities firms. These clearing members stood between the clearinghouse and the exchange's clients — firms like Tsingshan, which traded through clearing members rather than directly with the clearinghouse. The structure meant that Tsingshan's margin obligations were, in the first instance, the obligations of the clearing member through which Tsingshan traded. If Tsingshan could not meet a margin call, the clearing member was responsible, and if the clearing member could not meet the obligation, the clearinghouse was responsible. The chain of responsibility, in a crisis, could produce a cascade: a default by Tsingshan would impose losses on its clearing member, which could impose losses on the clearinghouse, which could impose losses on the other clearing members, which could impose losses on their clients. The clearinghouse had been designed to prevent exactly this kind of cascade, by calling margin in time to close out positions before they became systemic. On the morning of March 8, the clearinghouse was struggling, and failing, to prevent it.

* * *

The London Metal Exchange, at approximately eight o'clock in the morning London time on March 8, 2022, suspended trading in nickel. The suspension was, by itself, an extraordinary step — the LME had not suspended trading in a metal for a prolonged period in its recent history. But the suspension was not, by itself, the most extraordinary step the exchange took that morning. The most extraordinary step came shortly after the suspension, when the exchange announced that it was cancelling all nickel trades executed after midnight London time on March 8.

The cancellation affected roughly four billion dollars of trades. Every purchase and sale of nickel executed on the LME's electronic trading system between midnight and eight in the morning — trades at forty-eight thousand dollars, at sixty thousand, at eighty thousand, at one hundred thousand dollars per tonne — was retroactively voided. The trades would be treated as if they had never happened. The buyers who had purchased nickel at what they believed were market prices would not receive the nickel. The sellers who had sold nickel at what they believed were market prices would not receive the payment. The profits and losses that the participants believed they had made, over the course of the morning, would be erased. The exchange had declared, in substance, that its own price-discovery mechanism had produced prices that were invalid, and that the trades executed at those prices were therefore invalid.

The decision was unprecedented. The LME had, in its century-and-a-half history, suspended trading on various occasions — in response to wars, to supply disruptions, to market stress. It had not, in its recent history, cancelled trades after the fact. The cancellation was, in the language of market structure, a "bust" — a retroactive voiding of executed trades. Busts are, in the normal operation of an exchange, rare and limited in scope. An exchange might bust a single trade, or a small number of trades, if the trades were executed in error — if a trader had entered the wrong price, or the wrong quantity, or the wrong contract. The LME's bust, on March 8, was not limited. It covered every trade, in every contract, executed over an eight-hour period. It was a declaration that the market had not been functioning as a market for the entire period.

The reason the exchange gave for the cancellation was the protection of the clearinghouse. If the trades had stood, the participants who had sold nickel at forty-eight thousand dollars and bought it back at one hundred thousand dollars would have generated enormous losses. The participants who had bought nickel at forty-eight thousand and sold it at one hundred thousand would have generated enormous profits. The losses, in the first instance, would have fallen on Tsingshan, which held the massive short position. Tsingshan, unable to meet the margin calls, would have defaulted. The default would have imposed losses on Tsingshan's clearing member, which would have imposed losses on the clearinghouse, which would have imposed losses on the other clearing members. The cascade, the exchange concluded, would have threatened the solvency of the clearinghouse, and the solvency of the clearinghouse was, in the architecture of the exchange, the thing that could not be allowed to fail. If the

clearinghouse failed, every participant in the market would have taken losses, and the market itself — the LME, as an institution — would have been at risk.

The cancellation was, in this framing, a choice between two bad outcomes. The first bad outcome was the failure of the clearinghouse, which would have imposed losses on every participant in the market and which would have called into question the integrity of the LME as a venue for metals trading. The second bad outcome was the retroactive voiding of executed trades, which would impose losses on the participants who had made profits on the morning's trades and which would call into question the integrity of the LME as a venue in which trades, once executed, could be relied upon. The exchange chose the second bad outcome, on the reasoning that the first bad outcome was worse — that the failure of the clearinghouse would have been catastrophic, while the cancellation of trades, however damaging to the exchange's reputation, was a damage that the exchange could survive.

The participants whose trades were cancelled were not, as one might imagine, pleased. Jane Street Global Trading, a large proprietary trading firm, had sold nickel at prices that would have produced roughly one hundred and fifty-three million dollars in profit. Elliott Management, the hedge fund run by Paul Singer, had sold nickel at prices that would have produced roughly four hundred and fifty-six million dollars in profit. Both firms, and several others, had executed trades on the LME's electronic system during the cancelled period, at prices that the exchange had published and that the firms had relied upon. The cancellation turned those profits into zeros. The firms, in the months and years that followed, sued the LME, arguing that the exchange had exceeded its authority, that the cancellation had destroyed their legitimate property, and that the LME, as an exchange, could not be trusted to honor its own trades.

The lawsuits were, in their legal arguments, substantial. The LME's rulebook did grant the exchange emergency powers to cancel trades in extraordinary circumstances, but the scope of those powers had never been tested in a cancellation of this magnitude. The plaintiffs argued that the cancellation was not, in fact, necessary — that the clearinghouse could have been protected by other means, such as a capital injection from the exchange's parent company, HKEX, or by a standstill arrangement with Tsingshan and the clearing members. The LME argued that the cancellation was necessary, that the alternatives would not have been sufficient, and that the exchange's emergency powers were broad enough to cover the decision.

The UK High Court, in a judgment issued in September 2023, sided with the LME. The court found that the exchange had acted within its powers, that the decision to cancel had been a legitimate exercise of the exchange's emergency authority, and that the cancellation had been, in the circumstances, a reasonable response to a genuine crisis. The judgment was a vindication for the LME — a judicial confirmation that the exchange's actions, however damaging to the plaintiffs, had been within the bounds of its authority. The judgment was a disappointment for the plaintiffs — a judicial determination that their profits, however legitimately earned, had been lawfully erased. The plaintiffs appealed. The appeals, at the time of writing, are ongoing, and the ultimate legal resolution of the nickel crisis may not be known for several years.

The lawsuits, regardless of their outcome, have done lasting damage to the LME's reputation among the sophisticated trading firms that constitute a substantial portion of its electronic trading

volume. Jane Street and Elliott Management are, by any measure, among the most sophisticated participants in the global financial markets. Their decision to trade on the LME was a decision to trust the exchange — to trust that the prices it published were reliable, that the trades it executed were binding, and that the exchange would honor its obligations as a market operator. The cancellation, and the subsequent litigation, have shaken that trust. The firms have not, as of the time of writing, withdrawn from the LME entirely — the exchange remains the primary venue for industrial metals trading, and the firms' clients continue to require access to LME prices and LME contracts. But the firms' willingness to provide liquidity to the LME's electronic market — to be the counterparties that buy and sell on the exchange's electronic system, providing the volume that makes the market function — has, by all accounts, been reduced. The reduction in liquidity has, in turn, made the LME's prices less reliable, in a circular dynamic that the exchange's management has been working, since the crisis, to reverse.

The cancellation, regardless of its legal outcome, has had lasting consequences for the LME's reputation. The exchange's principal product — the thing it sells to its participants, and the thing that the global metals industry relies upon — is reliable price discovery. The LME's prices are used, around the world, to value physical metal contracts, to settle derivative positions, and to make investment decisions. The reliability of the prices depends on the reliability of the trades that produce them. If the trades can be cancelled, retroactively, by the exchange, then the prices that the trades produce are not reliable — they are provisional, subject to the exchange's subsequent determination that the market was functioning properly. The cancellation, in this sense, did more than void four billion dollars of trades. It called into question the integrity of every price the LME had ever produced, and every price it would produce in the future. The participants in the market could no longer assume that a trade executed on the LME was, in fact, a trade. They had to assume, instead, that a trade executed on the LME was a trade unless and until the exchange decided, retrospectively, that it was not.

The week that followed the cancellation — the week during which the LME kept nickel trading closed, from March 8 to March 15 — was, for the global metals industry, a period of profound uncertainty. The LME's nickel price was the benchmark against which physical nickel contracts around the world were settled. Producers who sold nickel to consumers, and consumers who bought nickel from producers, used the LME price as the reference for the price they negotiated in their physical contracts. With the LME price absent — with no official price being set, because no trading was occurring — the physical market had no benchmark. Producers and consumers were, in effect, flying blind. They negotiated prices bilaterally, or they used the last available LME price before the suspension, or they estimated what the price would be if trading resumed. The uncertainty was substantial for an industry that depended on the LME's price for its daily operations.

The negotiations that took place during the week of suspension, between the LME, Tsingshan, and the clearing members, were conducted behind closed doors, with limited public information. The substance of the negotiations, as it subsequently emerged, was a standstill agreement under which Tsingshan would meet its margin obligations over time, rather than immediately, and under which the clearing members and the exchange would provide the liquidity necessary to keep the clearinghouse solvent while Tsingshan's position was reduced. The standstill agreement allowed Tsingshan to avoid a

formal default, and allowed the clearinghouse to avoid the cascade of losses that a default would have triggered. The arrangement was a private-sector rescue, brokered by the exchange, in which the participants in the market agreed to absorb the cost of preventing a failure that would have imposed losses on all of them. The cost of prevention was, in the end, less than the cost of the alternative.

When the LME reopened nickel trading on March 16, it did so under new rules. The exchange imposed price limits — a maximum percentage move, up or down, from the previous day's settlement price, with an automatic suspension if the limits were breached. The limits were set at five percent for the initial reopening, and were subsequently adjusted. The exchange also announced a review of its clearing procedures, its margining methodology, and its emergency powers. The review, which was conducted over the following months, produced a series of reforms: more frequent margin calls, greater transparency over large positions, and a reconsideration of the exchange's ability to handle positions that, like Tsingshan's, were large relative to the deliverable supply.

The reforms were, in their specifics, sensible. They addressed the specific failures that the March crisis had exposed — the inadequate margining for large positions, the absence of effective price limits, the lack of transparency over the concentration of positions in the market. Whether the reforms will be sufficient to prevent the next crisis is, as with all regulatory reforms, an open question. The reforms are designed for the conditions that produced the last crisis. The next crisis, when it comes, will likely be produced by conditions that the reforms did not anticipate — a different metal, a different participant, a different kind of position, a different trigger. The reforms will reduce the probability of a repeat of the specific events of March 2022. They will not eliminate the underlying vulnerability, which is the vulnerability of a clearing system designed for gradual price moves to a market in which prices can move, under stress, at a speed that the clearing system cannot match.

The LME nickel crisis was not the first time that a modern commodities market had produced an outcome that its infrastructure was not designed to handle. In April 2020, the West Texas Intermediate crude oil futures contract — the benchmark for U.S. oil prices, traded on the New York Mercantile Exchange — had settled at a negative price, minus thirty-seven dollars per barrel. The negative price was the product of a specific set of circumstances: the COVID-19 pandemic had caused global oil demand to collapse, the world's oil storage facilities were approaching capacity, and the holders of WTI futures contracts, facing physical delivery of oil they had no place to store, were willing to pay to get rid of the contracts. The exchange's systems, which had been designed to handle positive prices, had not been designed to handle negative ones. The settlement at minus thirty-seven dollars produced a series of operational problems — margin calculations that broke, risk models that failed, retail investors who held oil futures through exchange-traded funds and suddenly found themselves owing money rather than holding an asset. The episode was, like the LME nickel crisis, a case in which the infrastructure of a commodities market encountered a situation it had not been built for, and in which the encounter produced outcomes that the market's participants had not anticipated.

The WTI negative prices and the LME nickel cancellation are, together, evidence of a broader pattern. The commodities markets' infrastructure — the exchanges, the clearinghouses, the margining systems, the price-discovery mechanisms — was built for a world in which commodities were produced, traded, and consumed by industrial participants, at a pace determined by industrial logistics.

In that world, prices moved gradually, positions were sized to physical needs, and the participants in the market were companies that produced or consumed the physical commodity. That world still exists, but it now coexists with a different world — a world in which commodities are traded by financial participants, at electronic speed, with positions sized to financial strategies rather than to physical needs. The two worlds share the same infrastructure, and the infrastructure, which was built for the first world, periodically fails when the second world produces conditions the infrastructure was not designed to handle. The failures will continue, because the infrastructure cannot be redesigned, in real time, to accommodate conditions that the infrastructure's designers did not anticipate. The failures will be different, each time, because the conditions that produce them will be different. The failures will share a common cause, which is the cause this book has been tracing: the rate at which the market can move has exceeded the rate at which the institutions that manage the market can respond.

The institutions that manage the world's commodities markets — the exchanges, the clearinghouses, the margining systems — were designed for a world in which prices moved gradually, in which positions were sized relative to deliverable supply, and in which the participants in the market were known to each other and could be relied upon to meet their obligations. That world existed, in its most fully developed form, for the century and a half of the LME's operation, from 1877 to roughly the turn of the twenty-first century. The world that has replaced it is one in which prices can move two hundred and fifty percent in two days, in which positions can exceed the deliverable supply, and in which the participants — hedge funds, proprietary trading firms, industrial companies operating from offices in Wenzhou — can hold positions that are too large for the market to absorb. The institutions have not adapted to the new world, and they may not be able to adapt, because the adaptation required — the construction of a clearing system that can handle, in real time, the price moves and position sizes that modern markets produce — is an undertaking that has no precedent in the history of commodities trading.

The LME's nickel crisis, in the end, was a crisis of speed. The price moved faster than the clearinghouse could call margin, the margin calls moved faster than Tsingshan could pay them, and the losses moved faster than the clearing members could absorb. The exchange's response — the suspension and the cancellation — operated on a human timescale: a decision made by the exchange's management, after consultation with its clearing members and its regulators, to stop the market and to void the trades that the market had produced. The response was effective — the clearinghouse survived, Tsingshan survived, the market survived. The response was also damaging — the cancellation destroyed the profits of participants who had executed trades in good faith, and it called into question the reliability of the LME as a venue for price discovery. The trade-off — solvency over integrity — was, in the circumstances, perhaps unavoidable. It was also a trade-off that the LME's founders, trading copper and tin in a coffeehouse above a hat shop in 1877, would not have recognized. They built an exchange that discovered prices. The exchange, in March 2022, discovered that it could not.

The nickel market, in the months and years after the crisis, has continued to function. Prices have settled, Tsingshan has reduced its position, and the LME has implemented its reforms. The lawsuits continue, and the reputation of the exchange, among the participants whose trades were cancelled, has not recovered. The LME's nickel price is still the global benchmark, used by producers and consumers

around the world, because there is no adequate substitute. The benchmark is used, however, with a recognition that it could, under stress, be suspended, and that the trades that produce it could, under stress, be cancelled. The recognition is a permanent condition of their reliance on the benchmark. They use it because they must. They use it with the knowledge that it is not, in the end, entirely reliable. That knowledge is the legacy of March 8, 2022, and it will endure for as long as the LME remains the benchmark — and perhaps longer, for as long as the participants in the commodities markets remember that an exchange, under sufficient stress, will choose its own survival over the integrity of its trades.

The image that this chapter leaves is the image of the Ring — the circle of red leather benches, in a room in the City of London, where traders have shouted bids and offers at each other for a century and a half. The Ring is the last open-outcry trading floor in Europe. It is a relic of a world in which prices were discovered by human beings, at human speed, in a physical place. The world has moved on. The prices are now discovered, in the main, by computers, in data centers, at electronic speed. The Ring remains, as a formality and as a symbol, but the prices it produces are, in a crisis, subordinate to the electronic market's prices — and the electronic market, on the morning of March 8, produced prices that the exchange could not accept. The Ring, on that morning, was silent. The electronic market was suspended. The prices, for a week, did not exist. The market that had been discovering prices for a century and a half discovered, on that morning, that it could not discover them at the speed the market now required. The discovery was the most consequential the LME had ever made.

Chapter 8

Four Hours to Mexico City

On a Tuesday morning in the San Fernando Valley, a man in a rented sedan drove through a residential neighborhood, stopping at intervals at small businesses — a car wash on Victory Boulevard, a restaurant on Van Nuys Boulevard, a furniture store on Sherman Way. At each stop, he collected a bag. The bags contained cash — American currency, in denominations of twenties, fifties, and hundreds, bundled with rubber bands and stuffed into grocery bags or shoeboxes. The cash was the weekly revenue from the sale of cocaine, methamphetamine, and other drugs in the Los Angeles metropolitan area. The businesses where the courier collected the cash were not, in their visible operations, drug-related. The car wash washed cars. The restaurant served food. The furniture store sold furniture. The businesses were, however, owned or operated by individuals who had agreed to serve as collection points — to receive the cash that the street-level dealers generated, to hold it in their safes or their back rooms, and to hand it to the courier when he arrived.

The courier was one of several dozen who operated in the Los Angeles area, each responsible for a route of collection points. By the end of his Tuesday route, the courier would have collected somewhere between two hundred thousand and five hundred thousand dollars in cash, depending on the volume of drug sales in his territory during the preceding week. The cash would be delivered to a stash house — an ordinary-looking residential property, typically rented under a false name, where the cash would be counted, sorted, and held until the next stage of its journey. The stash house was a way station, not a destination. The cash would be picked up, within a day or two, by another courier, who would deliver it to a representative of a peso broker — a person who managed the interface between the cash in the United States and the financial system that would, without ever touching a bank, transform it into usable currency elsewhere.

The peso broker's representative, upon receiving the cash, would send a message — by encrypted messaging app, in the current iteration of the system, though earlier iterations used telephone, fax, or even written notes carried by courier — to the peso broker's office, which was typically located in Mexico City, Guadalajara, or Cali. The message would report the amount collected: five hundred thousand dollars, say, from the Los Angeles route. The peso broker, upon receiving the message, would record the amount in a ledger. The ledger was the heart of the operation. It was a clearinghouse — a record of obligations and availabilities, matched and settled through a process that did not require the physical movement of money across any border.

The Black Market Peso Exchange — the name given by the Drug Enforcement Administration to the system that has, since the late 1980s, been the principal mechanism for laundering Colombian and Mexican drug proceeds — is, despite its name, less a market than a clearing arrangement. The system matches two sets of participants who have complementary needs. The first set consists of drug trafficking organizations that have large quantities of United States dollars, in cash, in American cities, and that need to convert those dollars into pesos, in Colombia or Mexico, for use in their operations.

The second set consists of legitimate or quasi-legitimate importers — Colombian importers of electronics and appliances, Chinese importers of raw materials and consumer goods — who need United States dollars, in the United States, to pay for goods purchased from American suppliers, and who have pesos, in their home countries, with which to pay for those dollars. The peso broker sits between the two sets, matching the dollars available with the dollars needed, and taking a commission on each match.

The mechanics of a typical transaction are straightforward in their outline. A Colombian importer wishes to purchase five hundred thousand dollars' worth of electronics from a supplier in Los Angeles. The importer does not have a dollar account at a United States bank, or does not wish to use it, or wishes to avoid the fees and the documentation that a formal wire transfer would require. The importer contacts a peso broker in Cali or Bogotá, and agrees to pay the broker, in pesos, in Colombia, the peso equivalent of five hundred thousand dollars, plus a commission of one to three percent. The peso broker, who has access to drug dollars held by couriers in the United States, instructs his representative in Los Angeles to deliver five hundred thousand dollars in cash to the electronics supplier (or to the importer's designated agent in the United States). The electronics supplier receives the cash, ships the goods to Colombia, and the importer receives the goods. The peso broker, meanwhile, has received the pesos from the importer in Colombia, and has delivered those pesos to the drug trafficking organization that generated the original dollars. The drug trafficking organization receives pesos in Colombia, which it uses to pay farmers, processors, transporters, and enforcers. The importer receives dollars in the United States, which it uses to buy goods. The peso broker takes his commission. The transaction has cleared. No dollar has passed through a bank. No dollar has crossed a border in a form that any regulator can see. The entire process, from the collection of the cash in Los Angeles to the delivery of the pesos in Colombia, has taken hours.

The system's efficiency, relative to the legitimate alternatives, is the reason it exists. A drug trafficking organization that has five hundred thousand dollars in cash in Los Angeles and that wishes to use those funds in Colombia has, in principle, several options. It can deposit the cash in a United States bank, wire it to a Colombian bank, and withdraw it in Colombia. This option is, for a drug trafficking organization, not available, because the deposit of five hundred thousand dollars in cash at a United States bank would trigger a Currency Transaction Report under the Bank Secrecy Act, would invite scrutiny from the bank's compliance officers, and would likely result in the freezing of the account and the reporting of the transaction to the Financial Crimes Enforcement Network. The organization can smuggle the cash, in bulk, from the United States to Colombia — by ship, by plane, by courier. This option is available, and is used, but it is expensive (the cost of logistics, the risk of seizure, the cost of corrupting customs officials) and slow (the cash must be physically transported, which takes days or weeks). The organization can use a cryptocurrency, such as Bitcoin, to transfer the value. This option is available, and is used for smaller amounts, but it is, for large volumes, less efficient than the BMPE — the conversion of cash to Bitcoin and back to cash incurs fees and risks that the BMPE avoids, and the blockchain, while not centrally supervised, is a public ledger that can be, and has been, analyzed by law enforcement.

The BMPE, by contrast, charges one to three percent, settles in hours, and leaves no trace in any financial system that any regulator supervises. The cash is collected in Los Angeles, delivered to a supplier or agent in the United States, and the offsetting pesos are delivered in Colombia. The dollars never leave the United States. The pesos never leave Colombia. The transaction is a barter — dollars in the United States exchanged for pesos in Colombia, with the physical movement of money eliminated by the matching of complementary needs. The barter is superior to any formal alternative. The formal alternatives — bank wires, Western Union, bulk cash smuggling — each impose costs that the BMPE does not require, take time that the BMPE does not take, and leave traces that the BMPE does not leave. The BMPE is the most efficient system that has ever existed for moving large sums of money across borders without documentation. It is more efficient than the legitimate banking system. It is more efficient than the hawala networks of the Middle East and South Asia. It is more efficient, for large volumes, than cryptocurrency. The reason for its efficiency is simple: it does not comply with anything. No forms are filed. No identities are verified. No reports are generated. The system operates outside the regulatory architecture, and the absence of the regulatory architecture is the source of its speed.

The history of the BMPE's development is worth sketching, because it illustrates the way that criminal systems, like legitimate ones, respond to the incentives created by regulation. Before the Bank Secrecy Act and the subsequent AML regime, the Colombian cartels of the 1970s and 1980s laundered their dollar proceeds in the simplest way available: they deposited the cash in American banks. The deposits were large — sometimes millions of dollars in a single account — and the banks, in the era before comprehensive AML requirements, accepted them without excessive scrutiny. The cash was then wired to Colombian banks, where it was withdrawn in pesos. The system was cheap and convenient for the cartels. It was also easy to trace — the wire transfers left records, the bank accounts were identifiable, and the pattern of deposits and withdrawals was, once identified, straightforward to prosecute.

The Bank Secrecy Act, and the subsequent AML regime, changed the calculus. The reporting requirements, the due diligence obligations, and the enhanced scrutiny of large cash deposits made it more difficult for the cartels to use the banking system directly. The cartels responded, initially, by structuring their deposits — breaking large amounts into many small ones, each below the reporting threshold, deposited at different banks by different individuals. The structuring technique, known as "smurfing," was effective for a time, but it was labor-intensive and vulnerable to detection — the patterns of small deposits, once identified by bank compliance officers or by FinCEN's analytical systems, could be traced back to the organizing network. The cartels needed a system that did not use banks at all. The BMPE, which had existed in nascent form since the 1970s, provided it.

The BMPE's growth in the 1990s was, in this sense, a direct consequence of the AML regime's effectiveness in closing the banking door. The regime made the banks safer. The criminals moved outside the banks. The regime, which was designed to supervise the banks, could not follow. The growth of the BMPE was not a failure of the AML regime, in the narrow sense — the regime had succeeded in making the banking system less useful to money launderers. The growth was, rather, a consequence of the regime's design, which assumed that money laundering would continue to pass through the banking system, and which did not contemplate the possibility that the criminals would

build a parallel system that bypassed the banks entirely. The assumption was reasonable when it was made. It was wrong, and the wrongness produced the BMPE.

* * *

The anti-money-laundering regime that the United States and its allies have built, over the past fifty years, is one of the most extensive and most expensive regulatory systems in the history of finance. It is also, by the measure of its effect on the specific criminal activity it was designed to prevent, one of the least effective. The disjunction between the regime's extent and its effectiveness is the subject of this chapter, and the disjunction is, like the other disjunctions this book has described, a product of the mismatch between the speed at which the regulated system operates and the speed at which the unregulated system has evolved.

The modern AML regime began with the Bank Secrecy Act of 1970. The act, passed in the closing years of the Nixon administration, was a response to the growing problem of organized crime's use of the financial system to launder illicit proceeds. The act's principal mechanism was the Currency Transaction Report — a form that banks were required to file with the Treasury Department for every cash transaction exceeding ten thousand dollars. The threshold was chosen, by the drafters of the legislation, as a level that would capture the large cash deposits associated with organized crime while avoiding the burden of reporting the ordinary cash transactions of legitimate businesses. The CTR was a dragnet: the Treasury would collect the reports, would maintain a database of large cash transactions, and would use the database to identify patterns that suggested money laundering.

The act was, for the first two decades of its existence, only moderately effective. The Treasury collected the reports, but the volume was large, the database was difficult to search, and the analytical resources available to the Treasury were limited. The drug trafficking organizations of the 1970s and 1980s — the Medellín and Cali cartels, principally — generated billions of dollars in cash from the sale of cocaine in the United States, and the initial laundering method was the simplest one: deposit the cash in American banks, in amounts just below the ten-thousand-dollar reporting threshold, using networks of individuals (known as "smurfs") who would visit multiple banks, on the same day, depositing nine thousand dollars at each. The technique, called "structuring," was a direct response to the Bank Secrecy Act's threshold. The threshold had been designed to catch large transactions. The criminals, by breaking their large transactions into many small ones, avoided the threshold. The law, in this sense, taught the criminals how to evade it. The ten-thousand-dollar reporting threshold became, for the criminal economy, not a barrier but a tutorial.

The Patriot Act, passed in the aftermath of the September 11 attacks, expanded the AML regime substantially. The act required banks to implement formal anti-money-laundering programs, to conduct customer due diligence, to file Suspicious Activity Reports for transactions that appeared suspicious regardless of amount, and to apply enhanced scrutiny to correspondent banking relationships with

foreign banks. The act also created the Financial Crimes Enforcement Network — FinCEN — as the Treasury's dedicated bureau for the collection and analysis of financial intelligence. The regime that emerged from the Patriot Act was, by any measure, extensive. Banks employed armies of compliance officers, who reviewed transactions, filed reports, and conducted investigations. The volume of reporting grew, year by year, until banks were filing millions of CTRs and SARs annually. The cost of compliance, borne by the banking industry and ultimately by its customers, grew into the tens of billions of dollars per year.

The regime's architecture was built on a single, fundamental assumption: that money laundering passes through the banking system. The assumption was, at the time the regime was designed, reasonable. The Medellín and Cali cartels of the 1980s and early 1990s did, in fact, launder much of their money through banks — through deposits, through wire transfers, through correspondent banking relationships with foreign institutions. The regime was designed to detect this activity, and it did, in fact, detect some of it. The reporting requirements, the due diligence obligations, and the enhanced scrutiny of correspondent banking relationships made it more difficult for criminal organizations to use the banking system in the ways they had used it in the 1970s and 1980s. The regime closed the door that the criminals had been using. The criminals, in response, walked through a different door.

The different door was the BMPE. The system, which had existed in nascent form since the 1970s, grew rapidly in the 1990s and 2000s, as the AML regime made it more difficult for criminal organizations to use the banking system. The growth of the BMPE was, in this sense, a direct response to the AML regime's success in closing the banking door. The regime had made the banks safer. The criminals had moved outside the banks. The regime, which was designed to supervise the banks, could not follow them.

The mechanism by which the BMPE evades the AML regime is structural, not tactical. The AML regime requires banks to file reports, to verify identities, and to monitor transactions. The BMPE does not use banks. The cash that the couriers collect in Los Angeles is delivered, in cash, to the supplier or agent who needs dollars in the United States. The pesos that the importer pays in Colombia are delivered, in cash or through domestic transfer, to the drug trafficking organization. No dollar crosses a border in a form that triggers a report. No identity is verified, because no bank is involved. No transaction is monitored, because no bank's systems are engaged. The AML regime, which is designed to detect transactions that pass through banks, is blind to transactions that do not pass through banks. The BMPE does not pass through banks. The BMPE is invisible to the AML regime.

The invisibility is not total. The Drug Enforcement Administration, the Federal Bureau of Investigation, and the Treasury's Office of Foreign Assets Control have, over the past three decades, conducted extensive investigations of the BMPE, and have achieved notable successes. The DEA's operations — with names like Operation Payload, Operation Cisco, and Operation■ — have dismantled specific peso broker networks, seized hundreds of millions of dollars in cash, and arrested dozens of individuals. The Treasury's designation of specific peso brokers as Significant Foreign Narcotics Traffickers, under the Kingpin Act, has frozen assets and imposed sanctions. The successes are real, and they have, in specific cases, disrupted the operations of specific organizations.

The successes have not, however, reduced the overall volume of money laundered through the BMPE. The system's volume has, by the estimates of the DEA and the Treasury, remained roughly stable at thirty to forty billion dollars per year, despite three decades of enforcement. The stability is not a consequence of enforcement's failure, in the conventional sense. The enforcement is competent and persistent. The stability is a consequence of the system's resilience — its capacity to replace the networks that are dismantled with new networks that serve the same function. The peso broker who is arrested is replaced, within weeks, by another peso broker who takes over his clients. The courier who is seized is replaced by another courier who takes over his route. The stash house that is raided is replaced by another stash house in another neighborhood. The system, like the internet, routes around damage. The damage, in the form of arrests and seizures, is absorbed. The system continues.

The reasons for the system's resilience are the same reasons for its efficiency. The system is decentralized — no single peso broker controls the market, and the failure of any single broker does not affect the overall capacity of the system. The system is demand-driven — as long as there are drug trafficking organizations with dollars in the United States and importers with pesos in Colombia, the demand for the BMPE's services will persist. The system is self-sustaining — the commissions that the peso brokers earn are sufficient to attract new participants to replace those who are removed by enforcement. The system is a parallel payments infrastructure, operating, like the M-Pesa system in Kenya, outside the formal banking system. The difference is that M-Pesa operates with the regulator's acquiescence, while the BMPE operates in defiance of the regulator's prohibition. The structural similarity is striking. The BMPE is the M-Pesa of the criminal economy. It is the payments system that the criminal economy built, in the space the formal system could not reach, because the formal system's compliance costs made it inaccessible.

The Chinese underground banking component, which merged with the BMPE in the 2000s, deserves separate treatment, because it illustrates the system's capacity to absorb new participants and new flows. The Chinese underground banking system — known in Chinese as fei ch'ien, or "flying money" — is a remittance system that has existed in various forms for centuries, originally developed by Chinese merchants to transfer funds between provinces without the physical movement of cash. The system operates through a network of brokers, typically based in Chinese communities in major cities, who accept deposits in one location and arrange for the equivalent amount to be made available in another location, through a matching of complementary flows — the same principle on which the BMPE operates. The Chinese system and the Colombian system, which had developed independently, converged in the 2000s, as Chinese importers in Shenzhen and Guangzhou began using peso brokers to obtain dollars for the purchase of goods from American suppliers. The convergence created a single, integrated system that could move money between the United States, Latin America, and China, without passing through any bank.

The integration has, in recent years, taken on a new dimension with the fentanyl trade. Fentanyl — the synthetic opioid that has, since the late 2010s, become the dominant cause of overdose deaths in the United States — is manufactured in Mexico, using precursor chemicals imported from China. The payment for the precursor chemicals flows through the same underground banking networks that the BMPE uses for drug proceeds. A Chinese chemical exporter ships precursors to a Mexican laboratory.

The Mexican laboratory, which is affiliated with a drug trafficking organization, pays for the precursors by directing its dollar holdings in the United States to the Chinese exporter's representative in the United States, through a peso broker or a Chinese underground banker. The Chinese exporter receives dollars in the United States, which it uses to purchase goods or to invest. The Mexican laboratory receives precursors in Mexico, which it uses to manufacture fentanyl. The transaction has cleared, again, without passing through any bank, and without triggering any report. The BMPE, which began as a system for laundering cocaine proceeds, has evolved into a system for financing the global trade in synthetic drugs. The evolution has occurred without any change in the system's fundamental architecture. The architecture is, as it has always been, a matching of complementary flows, settled outside the formal financial system, at a speed and a cost that the formal system cannot match.

* * *

The Black Market Peso Exchange is, by the metric that matters most to the people who use it, the most efficient payments system in the world. It moves thirty to forty billion dollars per year across borders, at a cost of one to three percent, in a matter of hours, without filing a single report, without verifying a single identity, and without passing through a single bank. No legitimate payments system — not SWIFT, not Visa, not Western Union, not the correspondent banking network — can match this combination of cost, speed, and reach. The legitimate systems are slower, more expensive, and more heavily documented. The illegitimate system is faster, cheaper, and undocumented. The difference is compliance. Compliance is a tax on speed and cost. The BMPE does not pay the tax. The legitimate systems do.

The comparison is concrete. A legitimate business that wishes to transfer five hundred thousand dollars from Los Angeles to a supplier in Colombia must, in the ordinary course, execute a wire transfer through its bank. The bank will verify the identity of the sender (under the Know Your Customer requirements of the Patriot Act), will screen the transaction against sanctions lists (under the Office of Foreign Assets Control's requirements), will assess whether the transaction appears suspicious (under the Suspicious Activity Report requirements), and will execute the transfer through the correspondent banking network — a chain of banks, each of which charges a fee and each of which takes time to process the transaction. The total cost of the transfer, in fees and in the time value of the funds during the settlement period, is typically between one and three percent. The total time, from initiation to settlement, is typically one to three business days. The transaction generates a paper trail — the wire transfer record, the correspondent bank records, the receiving bank's records — that is available to regulators and to law enforcement.

The BMPE, for the same transfer, charges one to three percent and takes four hours. The transaction generates no paper trail. The sender's identity is not verified. The transaction is not screened against sanctions lists. No report is filed. The money arrives, in the form of pesos in Colombia, before

the legitimate wire transfer would have cleared its first correspondent bank.

The efficiency advantage of the BMPE is not a function of superior technology. The BMPE uses the same technology that the legitimate system uses — telephones, encrypted messaging apps, spreadsheets. The efficiency advantage is a function of the absence of compliance. The BMPE does not verify identities because it has no regulatory obligation to do so. It does not screen against sanctions lists because it has no regulatory obligation to do so. It does not file reports because it has no regulatory obligation to do so. It does not use correspondent banks because it does not use banks. The absence of these obligations is the source of the BMPE's speed. The obligations, which the legitimate system bears, are the reason the legitimate system is slower.

This is the irony of the AML regime, and it is the irony that this chapter is concerned with. The AML regime, which was designed to prevent money laundering, has made the legitimate financial system slower and more expensive than the illegitimate system. The regime has, in effect, taxed the legitimate system for the privilege of compliance, and has allowed the illegitimate system to operate free of the tax. The result is that the illegitimate system is more competitive than the legitimate system for moving large sums of money across borders. The criminals are winning the efficiency race, not because they are smarter, but because they are not required to carry the burden that the legitimate system carries.

The comparison to other payments systems discussed in this book is instructive. M-Pesa, the Kenyan mobile money system, achieved its efficiency by operating outside the banking system — by building a payments rail that did not require a bank account, that did not charge banking fees, and that settled in seconds rather than days. The Central Bank of Kenya allowed M-Pesa to grow because it recognized that the public interest in cheap, accessible payments outweighed the banks' interest in maintaining their fee structure. The BMPE achieves its efficiency by the same means — operating outside the banking system, building a payments rail that does not require a bank account, settling in hours rather than days. The difference is that the BMPE's operators are criminals, and the flows they are moving are the proceeds of drug trafficking. The structural similarity is striking. Both M-Pesa and the BMPE are responses to the same underlying condition: the formal banking system, burdened by compliance costs and by the economics of serving low-value or high-risk transactions, left a gap. The gap was filled. In Kenya, it was filled by a telecom, with the regulator's acquiescence. In the Americas, it was filled by criminal networks, in defiance of the regulator's prohibition. The gap, in both cases, was the same gap — the space between what the banking system could do and what the market needed it to do. The banking system, encumbered by its regulatory architecture, could not serve the need. The parallel system, unencumbered, could.

The comparison to Tether is also instructive. Tether achieved its growth by issuing dollar-denominated liabilities outside the banking system, without a banking license, without full audit, and without comprehensive supervision. Tether's efficiency, relative to the regulated banking system, is a function of the same absence that gives the BMPE its efficiency: the absence of compliance. Tether does not file the reports that banks file. Tether does not maintain the capital that banks maintain. Tether does not submit to the supervision that banks submit to. The absence makes Tether faster, cheaper, and more flexible than the banks. The absence also makes Tether more dangerous, in the same way that the

BMPE's absence makes it more dangerous — the absence of compliance is the absence of the safeguards that compliance provides. The safeguards, in both cases, were designed to protect the system. The system, in both cases, has been bypassed. The bypassing has produced efficiency. The efficiency has produced growth. The growth has produced a parallel system that the regulators cannot see, cannot supervise, and cannot control.

The BMPE, Tether, and M-Pesa are, in their structural similarity, three instantiations of the same underlying pattern. The formal financial system, designed in an era when money moved through banks, imposes costs on its participants that reflect the assumptions of that era. The assumptions — that money moves through banks, that banks can be supervised, that supervision prevents abuse — held for most of the twentieth century. They do not hold now. Money moves through phones, through blockchains, through peso brokers, through courier networks. The institutions that were designed to supervise the banks cannot supervise these alternative channels. The alternative channels, unburdened by the costs of supervision, grow. The formal system, burdened by the costs, shrinks, or at least fails to grow as fast. The gap between the formal system and the alternative channels widens. The gap is the patchwork. The patchwork is where the money goes.

The burden is substantial. The legitimate banking industry spends, by various estimates, tens of billions of dollars per year on AML compliance. The spending employs armies of compliance officers, who review transactions, file reports, conduct investigations, and maintain the systems that the AML regime requires. The spending also generates a volume of reporting beyond the capacity of any human being to review. The banks file millions of Currency Transaction Reports and Suspicious Activity Reports annually. FinCEN, which receives the reports, processes them through automated systems that flag patterns, but the patterns that the systems are designed to detect are patterns in the banking data — deposits, withdrawals, wire transfers, and other transactions that pass through the regulated system. The BMPE does not pass through the regulated system. The patterns that the BMPE generates — cash collections in Los Angeles, peso deliveries in Colombia, goods shipments from the United States to Latin America — are not, in their individual components, suspicious. A cash deposit at a car wash is not suspicious. A shipment of electronics from Los Angeles to Bogotá is not suspicious. A peso transfer between two Colombian companies is not suspicious. The BMPE's transactions are, individually, invisible. They become visible only when they are assembled into the full picture of the system — a picture that no single regulator, and no single bank, is in a position to assemble, because the components are distributed across multiple jurisdictions, multiple institutions, and multiple participants.

The DEA, which has primary responsibility for the investigation of drug trafficking organizations, has been investigating the BMPE for three decades. The investigations are painstaking and resource-intensive. A single investigation — involving surveillance of collection points, interception of communications, tracing of cash flows, and coordination with foreign law enforcement — can take years, and can result in the dismantling of a single peso broker's network. The dismantling is a temporary disruption. The network that is dismantled is, within weeks, replaced by another network that serves the same clients, uses the same routes, and operates on the same principles. The DEA's agents know this. The agents investigate, arrest, and prosecute, knowing that the system they are

investigating will continue to operate, at roughly the same volume, regardless of their success. The agents do their jobs because their jobs are to enforce the law, and the law prohibits the activity they are investigating. The agents also know that the law they are enforcing is designed around the banking system, and that the system they are investigating is not the banking system.

The institutions that were designed to manage the financial system — in this case, the AML regime, the banks, and the law enforcement agencies — were designed for a world in which money laundering passed through banks. The criminals, faced with a regime that closed the banking door, built a parallel system that did not use banks. The parallel system grew, and is now, by volume, the principal mechanism for laundering the proceeds of the drug trade. The institutions that were designed to prevent money laundering cannot see the parallel system, because the parallel system does not pass through the institutions they supervise. The institutions are guarding a door the criminals no longer use. The criminals have built another one, and it is faster, cheaper, and invisible.

The AML regime's response to the BMPE has been, over the past two decades, a series of efforts to extend the regime's reach beyond the banks. The Patriot Act's correspondent banking requirements were, in part, an effort to capture the foreign banks through which some BMPE transactions passed. The Treasury's designation of specific peso brokers under the Kingpin Act was an effort to impose sanctions on the system's participants. FinCEN's advisories on trade-based money laundering were an effort to alert banks to the patterns that the BMPE produces in the legitimate trade data. The efforts have been sensible. They have not, however, been sufficient, because the BMPE's fundamental architecture — the matching of complementary flows, settled outside the banking system — is not susceptible to a regulatory response that is designed around the banking system. The regime can extend its reach, but the extension is always one step behind the system it is trying to reach. The system adapts. The regime follows. And thirty to forty billion dollars flows, every year, through the space between them — from the streets of American cities to the operations of the drug trafficking organizations, through a system that no bank sees and no regulator supervises.

The irony is worth stating plainly. The United States government, through the Bank Secrecy Act and the Patriot Act, has imposed on its banking system a compliance burden that costs tens of billions of dollars per year, employs hundreds of thousands of people, and generates millions of reports. The compliance burden was designed to prevent money laundering. The money laundering has not been prevented. It has been displaced — from the banking system, which is now heavily monitored, to the BMPE, which is not monitored at all. The compliance burden has made the banking system safer, in the sense that the banking system is now less useful to money launderers. The compliance burden has also made the BMPE more competitive, in the sense that the BMPE, which does not bear the compliance burden, can offer lower costs and faster settlement than the banking system. The regime has, in effect, subsidized the competition. The competition has grown. The regime has not caught up.

The BMPE is a patchwork — a parallel payments rail that has grown up alongside the formal banking system, serving the participants whom the formal system cannot or will not serve. The patchwork, in this case, is criminal. The patchwork exists not because the legitimate system is technically incapable of moving the money, but because the legitimate system is required to document the movement, and the documentation is unacceptable to the participants. The participants — the drug

trafficking organizations, the importers, the Chinese chemical exporters — do not wish to document their transactions, because the documentation would expose them to law enforcement. The BMPE offers them what the legitimate system cannot: movement without documentation. The offer is attractive. The offer has been accepted, to the tune of thirty to forty billion dollars per year.

The question that the BMPE poses is whether the AML regime, as currently designed, can ever be effective against a payments system that does not use banks. The question is the same question that the Central Bank of Kenya faced and the dollar regulators faced. The institutions were designed for a world in which money moved through banks. The world has changed. Money now moves through phones, through blockchains, through peso brokers, through Chinese underground bankers, through courier networks that operate on the streets of American cities. The institutions have not changed. The institutions guard the banks. The money has left the banks. The institutions are guarding an empty room.

The fentanyl trade, which has, since the late 2010s, become the dominant driver of drug overdose deaths in the United States, illustrates the BMPE's capacity to absorb new flows and serve new criminal economies. Fentanyl is a synthetic opioid, manufactured in laboratories, primarily in Mexico, from precursor chemicals imported from China. The manufacture of fentanyl does not require the cultivation of poppy fields or the processing of opium, as the manufacture of heroin does. It requires, instead, the importation of chemical precursors — fentanyl analogs, pressing agents, and other chemicals — which are legally manufactured in China and which are shipped, by legitimate commercial carriers, to Mexico. The payment for the precursors flows through the same underground banking networks that the BMPE uses for drug proceeds. A Chinese chemical exporter ships precursors to a Mexican laboratory. The Mexican laboratory pays for the precursors by directing its dollar holdings in the United States to the Chinese exporter's representative in the United States, through a peso broker or a Chinese underground banker. The Chinese exporter receives dollars in the United States, which it uses to purchase goods or to invest. The Mexican laboratory receives precursors in Mexico, which it uses to manufacture fentanyl. The transaction has cleared, again, without passing through any bank, and without triggering any report.

The fentanyl dimension is the most troubling aspect of the BMPE's evolution. The system that was built to launder cocaine proceeds has, in the 2010s and 2020s, been repurposed to finance the trade in the drug that is killing more Americans than any other. The repurposing required no change in the system's architecture. The peso brokers, the couriers, the stash house operators, the Chinese underground bankers — all continued to perform the same functions they had always performed. The only change was in the nature of the goods that the system's payments were financing. The change, from cocaine to fentanyl, increased the lethality of the trade that the BMPE supported, without increasing the system's visibility to the institutions that were supposed to prevent it. The Department of Justice, in a series of indictments announced in 2022 and 2023, targeted Chinese underground banking operators connected to the fentanyl precursor trade. The indictments were substantial — they named specific individuals, detailed specific transactions, and seized specific assets. The indictments were also incremental. The networks that were dismantled will be replaced, as they have always been replaced, by other networks that serve the same function. The fentanyl precursors will continue to flow, the

payments will continue to clear, and the overdoses will continue to occur.

The BMPE's persistence, in the face of three decades of enforcement, is the clearest illustration of the thesis that this book has been developing. The institutions that were designed to manage the financial system — the banks, the regulators, the law enforcement agencies — were designed for a world in which money moved through identifiable channels, at a speed that the institutions could match, and through institutions that the institutions could supervise. The BMPE does not move through identifiable channels. It moves through cash collections on the streets of American cities, through encrypted messages, through handshakes and ledger entries in offices in Guadalajara. It moves in hours, while the institutions move in months or years. It moves through criminal networks that have no formal existence, no regulatory standing, and no supervisable architecture. The institutions cannot see it, cannot stop it, and can, at the margins, only disrupt specific networks within it. The disruption is, in the context of the overall system, a cost of doing business — a tax that the system absorbs without contracting.

The DEA agent, reviewing his case file in his field office, knows this. The agent has spent years investigating a single peso broker network. He has conducted surveillance, intercepted communications, traced cash flows, and coordinated with colleagues in Colombia and Mexico. He has built a case that will, when it is filed, result in the indictment of a dozen individuals, the seizure of several million dollars in cash, and the disruption of a network that has been laundering drug proceeds for a decade. The agent knows that the disruption will be temporary. He knows that the network will be replaced. He knows that the system — the BMPE, the parallel rail, the patchwork — will continue to operate, at roughly the same volume, after his case is filed and the indictments are unsealed. The agent files the case anyway, because filing cases is what the agent does, and because the law requires that cases be filed when the evidence supports them. The agent also knows, in a way that he may not articulate but that his experience has taught him, that the system he is investigating is a rebuke to the system he is defending. The legitimate system, with its tens of billions of dollars in compliance spending and its millions of reports, cannot match the speed or the cost of the illegitimate system. The illegitimate system, with its couriers and its encrypted messages and its ledger of complementary flows, can move money from Los Angeles to Colombia in four hours, at one percent, without filing a single form. It does this every day, thirty to forty billion dollars a year, and has been doing it for three decades. It will continue to do it, in all probability, for as long as the legitimate system bears the compliance costs that make the illegitimate system competitive.

The street in the San Fernando Valley, where the courier made his Tuesday collections, is, at this moment, quiet. The car wash is washing cars. The restaurant is serving lunch. The furniture store is selling furniture. The cash that passed through these businesses yesterday is, by now, in the hands of the peso broker's representative, who has reported the amount to the broker in Guadalajara. The broker has matched the dollars with an importer's peso obligation, and has instructed the delivery. The importer will receive the dollars in the United States. The drug trafficking organization will receive the pesos in Colombia. The goods will be shipped. The drugs will be manufactured. The overdoses will occur. The cash will be collected, again, next Tuesday, by another courier in another rented sedan, on another route through another set of small businesses in another American city. The system will continue, as it has

continued for three decades, operating in plain sight, on the streets of American cities, every day, at a speed and a cost that the legitimate system cannot match.

Chapter 9

The Public Rail

A street vendor in Mumbai sells tea for ten rupees, which is approximately twelve American cents. The tea is brewed in a large aluminum pot, over a charcoal burner, on a wooden cart that the vendor pushes, every morning, from his home in a nearby chawl to a street corner in the business district. The cart is a modest operation. The vendor's customers are office workers, autorickshaw drivers, and other vendors, who stop for a cup in the morning or the afternoon, pay their ten rupees, and continue about their business. The transaction is unremarkable. What is remarkable, to a visitor from a country with a more developed financial system, is the manner in which the ten rupees is paid.

Taped to the side of the cart, at roughly eye level, is a piece of cardboard, perhaps four inches by six, on which is printed a QR code. The code is a black-and-white square, of the kind that has become, in the past decade, ubiquitous on restaurant tables, shop counters, and product packaging around the world. The customer, having drunk the tea, opens an application on his mobile phone — one of several dozen available in India, of which the most popular are PhonePe, Google Pay, and Paytm. The application uses the phone's camera to scan the QR code. The code identifies the vendor's bank account, through a system called the Unified Payments Interface, or UPI. The customer enters the amount — ten rupees — and approves the payment by entering a personal identification number. The ten rupees is debited from the customer's bank account and credited to the vendor's bank account. The transaction settles in a few seconds. The vendor's phone, propped against the tea pot, emits a chime indicating that the payment has been received. The customer walks away. The vendor pours the next cup.

The transaction has cost neither the customer nor the vendor anything. There is no fee. There is no minimum purchase. There is no monthly subscription. There is no percentage taken by a card network, no per-transaction charge by a payment processor, no surcharge added to the price of the tea to cover the cost of accepting the payment. The ten rupees that the customer pays is the ten rupees that the vendor receives. The transaction has settled through a payments system that is operated by the National Payments Corporation of India — a non-profit organization owned by a consortium of Indian banks and regulated by the Reserve Bank of India — and that is, by the design of its architects, free at the point of use.

The contrast with the payments infrastructure available in the United States is, to anyone who has attempted to accept a small payment in both countries, conspicuous. An American small business that wishes to accept a payment of ten cents — the approximate American equivalent of the ten-rupee tea — through a credit card will pay, to the card network and the payment processor, a fee of roughly 2.9 percent of the transaction amount plus thirty cents. The fee, on a ten-cent transaction, is thirty cents and three-tenths of a cent, which rounds to thirty cents. The fee, in other words, exceeds the transaction. The business that accepts the ten-cent payment by credit card receives, net of fees, approximately minus twenty cents. This is not, in the American payments system, a viable transaction. The American payments system, which was designed for transactions of a size sufficient to make the thirty-cent fixed

fee a small fraction of the total, is not designed for transactions of ten cents. The Indian payments system, which was designed for transactions of any size, is. The Indian street vendor can accept a ten-cent payment. The American small business cannot. The reason is not that the Indian vendor is more sophisticated. The reason is that the Indian payments infrastructure was built differently.

UPI was launched in April 2016 by the National Payments Corporation of India, with twenty-one partner banks. The launch was modest. There were no billboards, no celebrity endorsements, no consumer marketing campaigns. The technology was the story: a unified, interoperable, real-time payments rail that connected any bank account to any other bank account, using a virtual payment address — a UPI ID, of the form name@bank — that required no account numbers, no branch codes, and no bank routing information. The transaction settled instantly, at no cost to the user. The system was open, in the sense that any bank that wished to participate could connect, and any technology company that wished to build a consumer-facing application on top of the system could do so. The system was public, in the sense that it was owned and operated by a non-profit consortium regulated by the central bank, rather than by a private company pursuing a profit. The system was the opposite of the Visa-Mastercard model that prevailed in the United States and Europe.

The Visa-Mastercard model, which has governed card payments in the developed world since the 1970s, is a private model. Visa and Mastercard are for-profit corporations, publicly traded, with market capitalizations in the hundreds of billions of dollars. They operate networks that connect banks to merchants, and they charge, for each transaction processed over their networks, a fee — the interchange fee, which is typically around two percent of the transaction amount, plus a fixed per-transaction charge. The fees are borne, in the first instance, by the merchants who accept the cards, and are passed on, in the second instance, to the consumers who buy the goods, in the form of higher prices. The system is a tax on consumption, levied by the card networks and distributed to the banks that issue the cards. The tax generates, for Visa and Mastercard and their member banks, revenues in the hundreds of billions of dollars annually. The tax is largely invisible — embedded in the price of goods, unitemized on receipts, and collected through a system that has become, over five decades, so entrenched that most consumers do not regard it as a tax at all.

The UPI model is the elimination of the tax. The National Payments Corporation of India does not charge a per-transaction fee. The banks that participate in UPI do not charge a per-transaction fee. The technology companies that build consumer applications on top of UPI — PhonePe, Google Pay, Paytm — do not charge a per-transaction fee. The transaction is free. The cost of operating the system is borne, not by the transactions themselves, but by the banks that own the NPCI, as a cost of maintaining their deposit franchises. The banks have concluded, in the Indian context, that the cost of providing free payments is less than the cost of losing deposits to competitors who provide free payments. The conclusion is one the American banks have not reached, because the American banks earn, from the Visa-Mastercard interchange system, revenues that the Indian banks do not earn. The American banks have no incentive to build a free payments rail, because the existing rail, which charges two percent, is profitable for them. The Indian banks, operating in a market where the Visa-Mastercard rail had not achieved the penetration it had achieved in the United States, had less to lose, and more to gain, by building a substitute.

The specifics of the UPI system's design illustrate the way that a public payments rail can be built to be both universal and free without, in the process, becoming a bureaucratic monstrosity. The system operates on a set of open APIs — application programming interfaces — that any bank or technology company can use to connect. The APIs specify the format of the payment messages, the authentication mechanisms, and the settlement procedures. The specifications are public, published by the NPCI, and updated through a process of consultation with the participating banks and technology companies. A bank that wishes to connect to UPI implements the APIs, integrates with the NPCI's central switching infrastructure, and begins sending and receiving payment messages. A technology company that wishes to build a consumer application on top of UPI partners with a bank (which provides the banking infrastructure and the regulatory compliance), implements the APIs, and launches its application. The consumer downloads the application, links it to a bank account, and begins making payments. The entire process, from the consumer's perspective, takes minutes. The process, from the bank's or the technology company's perspective, takes weeks or months, depending on the complexity of the integration. The process is, in either case, well-documented, standardized, and open.

The settlement mechanics are straightforward. When a customer initiates a payment through a UPI application, the application sends a message, through the NPCI's switching infrastructure, to the customer's bank. The bank verifies that the customer has sufficient funds, debits the customer's account, and sends a confirmation message back through the NPCI to the recipient's bank. The recipient's bank credits the recipient's account. The entire process takes, in normal operation, a few seconds. The settlement is real — the funds are actually moved between the banks, through the Reserve Bank of India's settlement accounts, not merely promised or posted. The customer sees the debit, the vendor sees the credit, and the transaction is complete. There is no batch processing, no overnight settlement, no waiting period. The ten rupees that the customer pays for the tea is, within seconds, in the vendor's bank account. The vendor can, if he wishes, immediately use those ten rupees to pay for his own purchases — to buy milk from a neighboring vendor, to pay an autorickshaw driver, to settle a debt with a friend. The money moves at the speed of the network, which is, in India in 2024, fast enough that the user does not notice the delay.

* * *

The story of UPI's development begins with a man named Nandan Nilekani. Nilekani is a figure of considerable standing in India. He was, in 1981, one of the seven co-founders of Infosys, the technology services company that became, over the following three decades, one of India's largest and most successful enterprises. Infosys was a pioneer of the Indian IT outsourcing industry, which grew, in the 1990s and 2000s, into a major component of the Indian economy. Nilekani's career at Infosys — he served as chief executive from 2002 to 2007, and as co-chairman thereafter — gave him both the financial resources and the institutional credibility to undertake projects of national scope. In 2009, the

Indian government appointed him to head the Unique Identification Authority of India, the body responsible for the Aadhaar project — the biometric identification system that would, over the following five years, enroll more than 1.2 billion Indian residents and issue each of them a unique, twelve-digit identification number linked to their biometric data (fingerprints and iris scans).

The Aadhaar project was an identity layer. Its purpose was to provide every Indian resident with a verifiable identity, in a country where hundreds of millions of people had, before Aadhaar, no formal identification documents and therefore no access to the formal financial system. A person without identification cannot, in most countries, open a bank account, obtain a mobile phone contract, or receive government benefits. The Aadhaar project was designed to solve this problem, by providing a universal, biometrically verified identity that any bank, telecom, or government agency could use to confirm that a person was who they claimed to be. The project was unprecedented in scale and ambition. No country had previously attempted to enroll more than a billion people in a biometric identification system. The project was controversial — concerns about privacy, about data security, and about the potential for surveillance were raised, and continue to be raised, by civil liberties organizations and by the Indian courts. The project had a substantial effect on financial inclusion. The number of Indians with bank accounts grew, in the years after Aadhaar's rollout, from roughly 35 percent of the adult population to over 75 percent.

Nilekani's role in the Aadhaar project gave him, in addition to the project's direct effects, a perspective on the architecture of India's digital infrastructure that few other people possessed. He had built the identity layer. The next layer, he recognized, was payments. The existing payments infrastructure in India, in the early 2010s, was a patchwork. Cash dominated, accounting for the vast majority of transactions by volume. Card penetration was low — Visa and Mastercard had achieved, in India, a fraction of the market share they held in the United States and Europe, because the merchants who would have accepted the cards found the interchange fees uneconomical for the small-ticket transactions that constituted the bulk of Indian retail. Mobile money systems, of the kind that had succeeded in Kenya, had not achieved comparable scale in India, because the Indian telecom market was fragmented and because the Reserve Bank of India had not adopted the regulatory forbearance that the Central Bank of Kenya had adopted. The patchwork was a constraint on the digital economy — a constraint that the identity layer, by itself, could not remove.

The National Payments Corporation of India, which had been established in 2008 by the Reserve Bank of India as a non-profit consortium of Indian banks, was the institution charged with building the payments infrastructure that the country lacked. The NPCI was a cooperative — owned by its member banks, regulated by the central bank, and operated as a public utility rather than as a for-profit enterprise. The NPCI had, before UPI, built several payment systems, including the Immediate Payment Service (IMPS), which allowed real-time interbank transfers, and the National Financial Switch, which connected the country's ATMs. These systems were technically adequate but not transformative. IMPS, in particular, required the sender to know the recipient's bank account number and the recipient's bank's IFSC code — a piece of information that most Indians did not have, for most of the people to whom they might wish to send money. The system worked, in the technical sense, but it was not accessible enough to drive mass adoption.

The UPI project, which the NPCI began developing in 2014 and launched in April 2016, was designed to address the accessibility problem. The key innovation was the virtual payment address — the UPI ID, of the form name@bank, that served as an alias for the user's bank account. A user who wished to send money to another user needed only the recipient's UPI ID, not the recipient's account number or bank code. The UPI ID was an email address for money — a simple, memorable identifier that could be shared as easily as a phone number or a social media handle. The innovation was technically modest. Its effect on adoption was substantial. The friction that had prevented IMPS from achieving mass adoption — the need to know and enter bank account numbers and IFSC codes — was eliminated. A user could send money to another user by entering a short, human-readable identifier. The transaction settled instantly. The cost was zero. The system was open to any bank and to any technology company that wished to build an application on top of it.

The system's adoption, in the months after its April 2016 launch, was modest. The transaction volumes were small. The partner banks were cautious. The technology companies that would later build the consumer applications that drove UPI's mass adoption — PhonePe, Google Pay, Paytm — had not yet launched their UPI-based products, or had launched them only in limited form. The system was, in the summer of 2016, a promising piece of infrastructure in search of a market.

The market arrived, with considerable abruptness, on November 8, 2016. On that evening, the Indian government announced, with four hours' notice, that all five-hundred-rupee and thousand-rupee currency notes — which together constituted roughly 86 percent of the cash in circulation in India — would cease to be legal tender at midnight. The population had until December 30 to deposit the old notes at banks, and would receive, in exchange, new notes that the government would print and distribute. The announcement, which came to be known as "demonetization," was an effort to flush out "black money" — untaxed cash holdings that the government alleged were financing criminal activity and corruption. The actual effects of demonetization, as they emerged over the following months and years, were complex, and the policy's effectiveness in achieving its stated purpose has been the subject of extensive and ongoing debate among economists.

The policy's effect on digital payments adoption, however, was unambiguous. The sudden voiding of 86 percent of the country's cash produced, in the weeks that followed, an acute shortage of the medium in which the vast majority of Indian transactions were conducted. The population, faced with the inability to pay in cash, sought alternatives. The alternatives were, in the immediate aftermath of demonetization, limited — the new notes were in short supply, the banks were overwhelmed by the volume of exchanges, and the electronic payment systems that existed were, for the most part, the systems that UPI had been designed to replace. UPI, which had been operational for seven months but had achieved only modest adoption, was positioned — by virtue of its accessibility, its instant settlement, and its zero cost — to absorb a portion of the demand for non-cash payments that demonetization had created. The absorption began within days of the announcement. The transaction volumes on UPI, which had been running at roughly one hundred thousand per month before demonetization, began to climb. By the end of 2016, the monthly volume had reached several million. By the end of 2017, it had reached hundreds of millions. By the end of 2019, it had exceeded one billion. By 2023, it was exceeding ten billion transactions per month — a volume that exceeded, by a

substantial margin, the combined monthly transaction volume of Visa and Mastercard globally.

The demonetization event illustrates it illustrates the way that institutional decisions — even decisions that are, by most economic assessments, poorly conceived — can produce unintended consequences that are substantial in their long-term effects. The stated purpose of demonetization was the elimination of "black money" — untaxed cash holdings that the government alleged were financing criminal activity and corruption. The actual effect of demonetization on black money was limited. The vast majority of the demonetized currency was, in fact, deposited in banks — which meant that the holders of the currency, whether they had acquired it legally or illegally, were able to exchange it for new notes. The policy did not, in any meaningful sense, "flush out" the black money that it was designed to flush out. The policy did, however, produce two effects that were not its stated purpose. The first effect was short-term economic disruption: the cash shortage reduced consumption, reduced economic activity, and imposed costs, particularly on the informal sector and on small businesses that operated primarily in cash. The second effect was the acceleration of digital payments adoption — an effect that was not the policy's goal, but that was, in retrospect, its most consequential and most lasting result. The acceleration occurred because the demonetization shock forced a large fraction of the Indian population, for the first time, to seek alternatives to cash. UPI was the alternative that was, by virtue of its launch seven months earlier, available. The availability, combined with the shock, produced the adoption. The adoption, once begun, continued — because the users who tried UPI during the demonetization crisis discovered that it was, in fact, more convenient than cash, and continued to use it after the crisis passed. The continuation is, in the long-term data, visible: UPI's transaction volumes did not decline after the demonetization crisis ended. They grew, and have continued to grow every year for the eight years since. The growth is now driven by preference, not by crisis — by the simple fact that the system, once tried, is more convenient than the alternative. The rail was built, the rail was free, and the population, once it had tried the rail, did not go back to cash.

The NPCI is not a conventional central bank. The NPCI is not the Reserve Bank of India. The NPCI is a non-profit corporation, owned by its member banks, regulated by the Reserve Bank, and operated as a utility. The arrangement is similar to the arrangement that governs SWIFT — a cooperative owned by its members, regulated by the relevant authorities, and operated as a utility. The difference is that the NPCI, unlike SWIFT, was directed by its regulator to build a specific piece of infrastructure — UPI — and was given the resources and the mandate to do so. The direction was effective, because the Reserve Bank of India has the authority to direct the NPCI, and because the NPCI, as a non-profit owned by the banks, has the incentive to comply. The arrangement has produced, in UPI, a piece of infrastructure technically comparable to the best the private sector has built anywhere, and economically superior to all of them — because it is free.

* * *

The growth of UPI, in the years after demonetization, was driven by a combination of factors that, taken together, produced an adoption curve steeper than any other payments system had achieved, in any market, in a comparable period. The factors included the demonetization shock, which created an immediate need for non-cash alternatives; the Aadhaar identity layer, which had, over the preceding five years, brought hundreds of millions of Indians into the formal financial system for the first time; the Jan Dhan financial inclusion program, a government initiative that had opened roughly four hundred and seventy million bank accounts for previously unbanked Indians; the smartphone penetration, which had grown, in the years before UPI's launch, to roughly five hundred million users; and the consumer applications — PhonePe, Google Pay, Paytm — that the technology companies built on top of the UPI rail, and that provided the user-friendly interfaces through which the mass of Indian consumers interacted with the system.

The consumer applications were nearly identical. Each provided a simple interface: the user opened the app, selected "send" or "receive," entered or displayed the relevant UPI ID or QR code, and confirmed the transaction with a PIN. The apps were distinct in branding — PhonePe was owned by Flipkart, the Indian e-commerce company that had been acquired by Walmart; Google Pay was owned by Google; Paytm was owned by One97 Communications, an Indian technology company. The apps were mechanically indistinguishable. All of them used the UPI rail. All of them settled transactions through the NPCI. All of them were, from the user's perspective, free. The competition among the apps was not a competition over payments technology. The competition was a competition over user acquisition, over merchant relationships, and over the auxiliary services — bill payment, mutual fund investment, insurance — that the apps offered on top of the basic payments function.

The competition produced a payments market that was remarkably efficient. The apps competed for users by offering better interfaces, faster onboarding, and more auxiliary services. They competed for merchants by offering easier QR code onboarding, better analytics, and integrated inventory management. They did not compete on price, because the price — zero — was set by the rail. The rail was public. The competition was on top of the rail. The arrangement was analogous to the arrangement that governs the internet itself — a public infrastructure (the TCP/IP protocol, the domain name system) over which private companies (Google, Amazon, Facebook) compete to provide services. UPI was, in this analogy, the TCP/IP of Indian payments. The apps were the websites.

The comparison to the American payments system is instructive. An American consumer who wishes to send money to another American consumer has, in principle, several options. The consumer can write a check, which the recipient must deposit at a bank, and which settles over a period of one to three business days through the check clearing system. The consumer can use a person-to-person payment service like Venmo or Zelle, which typically settles within one business day, and which may, depending on the funding source, charge a fee. The consumer can use a credit card, which charges the recipient a fee of two to three percent. The consumer can use cash, which requires physical proximity. The American system is slower and more expensive than the Indian system, because the American system is built on a rail — the ACH system, the card networks, the check clearing system — that was designed in an era when speed was less important and when the economics of payments were based on fees rather than on ubiquity.

The ACH system, which is the backbone of American electronic payments, was designed in the 1970s. It settles in batches, once per business day, which means that a payment initiated on a Friday does not settle until the following Tuesday. The system is slow. The Federal Reserve, which operates the ACH system, has, in recent years, introduced a real-time payment system called FedNow, which settles instantly and which is technically comparable to UPI. FedNow launched in July 2023. As of the time of writing, its adoption has been limited — a small fraction of American banks have connected, and the transaction volumes are negligible by comparison with UPI's. The reasons for the limited adoption are structural. The American banks, which earn substantial revenues from the existing payments infrastructure — from interchange fees, from overdraft fees, from the float on deposited funds — have little incentive to migrate their customers to a system that settles instantly and charges nothing. The banks' interest is in the preservation of the existing infrastructure, not in its replacement. The result is that the United States, in 2024, has a real-time payment system that is technically capable but commercially marginal, while India has a real-time payment system that is both technically capable and commercially dominant.

The reasons for the divergence are not, primarily, technological. The technology that UPI uses — the APIs, the real-time switching, the mobile applications — is technology that is available in every country, and that has been implemented, in various forms, by payments operators around the world. The reasons for the divergence are institutional. The Reserve Bank of India, in establishing the NPCI as a non-profit and in directing it to build a public rail, made a decision that the Federal Reserve has not made. The Federal Reserve operates the ACH system, and has now launched FedNow, but it has not, in the manner of the NPCI, built a public payments rail that is open to private-sector competition at the application layer and free at the point of use. The Federal Reserve's approach to payments infrastructure has been, historically, to provide the backbone — the ACH system, the wire transfer systems — while leaving the consumer-facing payments to the private sector. The private sector, in the United States, has built the Visa-Mastercard network, the Venmo and Zelle services, and a range of other payment products, all of which operate on top of the Federal Reserve's backbone, and all of which charge fees that are, by comparison with UPI's, substantial. The fees are, in the American context, accepted — by the merchants who pay them, by the consumers who bear them, and by the regulators who have not, so far, intervened to eliminate them. The fees are, in the Indian context, not accepted — because the NPCI, by building a free rail, has demonstrated that the fees are not necessary, and that a payments system can operate without them.

The demonstration has not been lost on the rest of the world. Central banks and payments regulators in Brazil, Thailand, Malaysia, and a number of other countries have, in the years since UPI's launch, built or launched their own real-time, low-cost payment systems — Pix in Brazil, PromptPay in Thailand, DuitNow in Malaysia. These systems are variations on the UPI model: public rails, open to private-sector competition, settling in real time, and charging little or nothing at the point of use. The systems have, in each country, achieved adoption curves that, while not as steep as UPI's, are substantially steeper than the adoption curves of the card networks in the same markets. The result is a global shift away from the private, fee-based payments model that has prevailed since the 1970s, and toward a public, free-at-the-point-of-use model that has been demonstrated, by India, to be viable at

scale. The shift has substantial implications for the card networks. The networks, which have built their businesses on the collection of interchange fees, are facing, in market after market, the emergence of a public alternative that does not charge interchange fees. The alternative is currently confined largely to domestic transactions within individual countries. The alternative is, however, growing, and the direction of the growth — more countries, more transactions, more value — is clear.

The Indian system's dominance is comprehensive. UPI processes, as of 2024, roughly fourteen billion transactions per month. The combined global transaction volume of Visa and Mastercard — across all card types, in all countries — is, by comparison, roughly ten billion per month. UPI, a system that did not exist in 2015, now processes more transactions than the two largest card networks in the world, combined. The comparison is somewhat inexact — the Visa and Mastercard figures include transactions of all sizes, in all currencies, across all merchant categories, while the UPI figures are domestic Indian transactions, predominantly small-ticket, in rupees. The comparison is instructive, because it illustrates the scale a public, zero-cost payments rail can achieve, in a market of India's size, when the rail is built and the private sector is allowed to compete on top of it. The scale has been achieved because the rail is free. The rail is free because the institution that operates it is a non-profit consortium, not a for-profit corporation. The institution is a non-profit consortium because the Reserve Bank of India, in 2008, decided to establish the NPCI as a non-profit, rather than licensing a private company to build the country's payments infrastructure.

The decision, in retrospect, appears prescient. At the time, it was unconventional. Most central banks, in most countries, had left the construction of payments infrastructure to the private sector — to the card networks, to the payment processors, to the banks. The Reserve Bank of India, by establishing the NPCI as a non-profit and by directing it to build a public rail, made a choice that most other central banks had not made. The choice was unremarkable in its immediate effects. The NPCI built its systems, the banks connected, and the transactions flowed. The choice was transformative, because the public rail that the NPCI built — UPI — provided the infrastructure on which the Indian digital payments economy was constructed. The economy that has grown on top of UPI — the consumer apps, the merchant services, the lending platforms, the insurance products — is more advanced than the equivalent economy in the United States, because the underlying rail is more efficient. The rail is more efficient because it was built to be a public utility, not a private toll road.

The Reserve Bank of India, in making the decision to build a public rail, was not acting out of a generalized hostility to private enterprise. The Indian government has, over the past three decades, pursued a broadly market-oriented economic policy, and the Indian private sector has, in the technology and financial services industries, grown substantially. The decision to build a public rail was, rather, a recognition that payments infrastructure is a natural monopoly — a piece of infrastructure that is most efficiently operated by a single provider, and that, if operated by a private provider, will generate monopoly rents that are borne by the users. The card networks, in the United States and Europe, are a duopoly — Visa and Mastercard — that generate, for their shareholders and their member banks, revenues that are, in economic substance, monopoly rents. The Reserve Bank of India, by building a public rail, eliminated the possibility of monopoly rents in the Indian payments market. The elimination has been a substantial benefit to the Indian consumer and the Indian merchant. It has been a substantial

loss to the card networks — a market of 1.4 billion people that has, in large part, bypassed the card networks entirely, and that will, in all probability, continue to bypass them.

The international expansion of UPI, which began in earnest in 2023, is the next chapter of the story. The NPCI has signed agreements with payments operators in the United Arab Emirates, Singapore, France, Sri Lanka, Mauritius, and Nepal, allowing Indian tourists to use UPI to make payments in those countries. The expansion is an effort to extend the UPI rail beyond India's borders, and to provide an alternative to the Visa-Mastercard duopoly in international payments. The expansion is currently limited — the acceptance network in the partner countries is small, and the transaction volumes are modest. The expansion is, however, a signal of intent. The NPCI, having built a payments rail that processes more transactions than Visa and Mastercard combined, is now exploring whether that rail can be extended to serve a global market. The exploration will take years, and the outcome is uncertain. The exploration is the most serious challenge to the Visa-Mastercard duopoly that has emerged in the fifty years since the duopoly was established. The challenge comes not from a competing private network, but from a public utility, built by a central bank, in a developing country, that has, in the space of eight years, achieved what no private network has achieved: a payments system that is simultaneously universal, instant, and free.

The challenge, if it materializes into a genuine global alternative, will have implications that extend beyond the payments industry. The Visa-Mastercard duopoly is an American-dominated system. The card networks are headquartered in the United States. The transactions processed over the networks are subject to United States law. The dollar-denominated transactions are subject to the jurisdiction of the United States Treasury and the Office of Foreign Assets Control. A global UPI alternative, if it achieved scale, would provide a payments rail that was not subject to American jurisdiction — a rail operated by an Indian non-profit, regulated by the Reserve Bank of India, and settled in rupees or in the currencies of the participating countries, rather than in dollars through the American banking system. The implications for the United States government's ability to use the payments system as an instrument of foreign policy would be substantial. The United States government, which has used the SWIFT exclusion as a sanctions weapon against Russia, and which has used the dollar clearing system as a sanctions weapon against Iran, would find that a growing fraction of the world's payments were being processed over a rail that it did not control and could not exclude. The infrastructure that was built in the twentieth century, under American auspices, is facing the emergence of alternatives that are built outside American auspices, and that are growing.

The Indian street vendor, pouring tea from his aluminum pot, is not, in all probability, aware of the institutional history that has made his QR code possible. He knows that the code works, that the money arrives, and that the system charges him nothing. He knows that the alternative — cash — requires him to carry change, to guard against theft, and to make deposits at a bank that is a bus ride away. The QR code is a convenience. The convenience is the product of a series of decisions — the decision to build Aadhaar, the decision to establish the NPCI, the decision to build UPI, the decision to demonetize, the decision to make the rail free — that were made, over a period of fifteen years, by a small number of people in positions of authority, none of whom the vendor has met, and all of whom the vendor has, in his daily use of the system, reason to be grateful for, whether he knows it or not. The gratitude is not the

point. The point is that the system works, that it works for free, and that it works for a ten-rupee transaction on a street corner in Mumbai, in a way that the equivalent system, in the United States, does not.

The vendor pours the tea. The customer drinks it. The ten rupees moves, in a few seconds, from one bank account to another, over a rail that did not exist ten years ago, and that now processes more transactions than the two largest payments networks in the world combined. The transaction is unremarkable. It is also the product of one of the most successful pieces of financial infrastructure ever built. The achievement is the ordinariness itself — that the system has become, for the vendor and his customer, as routine as cash. That ordinariness took eight years to build, and it will, in all probability, endure.

Chapter 10

Eleven Dollars

In the autumn of 1958, the Bank of America, which was at that time the largest bank in California and one of the largest in the United States, executed a marketing maneuver that has come to be known, in the relatively small literature devoted to the history of consumer credit, as the Fresno Drop. The maneuver consisted of mailing, unsolicited, approximately sixty thousand credit cards to residents of Fresno, California. The cards were functional. They could be used, immediately, at any of the merchants in Fresno who had agreed to participate in the Bank of America's credit card program. The cardholders had not applied for the cards. They had not been screened for creditworthiness. They had not consented to receive them. The cards arrived in their mailboxes, in envelopes, with a letter explaining that they could be used to make purchases at participating merchants, and that the charges would be billed monthly, with interest accruing on unpaid balances.

Fresno was chosen, by the Bank of America's marketing department, for reasons that were practical rather than romantic. Fresno was, in 1958, a city of approximately 130,000 people, located in the Central Valley of California, far enough from the major metropolitan centers of San Francisco and Los Angeles that the experiment, if it failed, would not produce a public embarrassment. The city's population was large enough to provide a meaningful test of the card's adoption, but small enough that the bank could manage the merchant acquisition and the customer service operations that the card's use would require. Fresno was, in the language of marketing, a test market — a place where a new product could be tried, and where the results could be observed, before the product was rolled out to the broader market.

The Fresno Drop was an extraordinary way to launch a financial product. The mailing of unsolicited credit cards to people who had not requested them, and who had not been assessed for their ability to repay the debts they might incur, would today be illegal — the Fair Credit Reporting Act of 1970, the Equal Credit Opportunity Act of 1974, and a series of subsequent regulations have made the practice impermissible. In 1958, the practice was legal, and the Bank of America, which was at that time operated by executives who were willing to take risks that later generations of bank executives would not take, undertook it. The risk was substantial. The cards could have been used fraudulently. The cardholders could have defaulted. The merchants could have refused to participate. The experiment could have failed, at substantial cost to the bank. The experiment did not fail. The cards were used. The merchants accepted them. The balances were paid, for the most part, and the interest on the unpaid balances generated revenue for the bank. The Fresno Drop was a success. The Bank of America, having validated the model in Fresno, expanded it to the rest of California, and then, through licensing agreements with banks in other states, to the rest of the United States.

The credit card that the Bank of America launched — the BankAmericard — was a revolution in consumer credit. Before the BankAmericard, consumer credit in the United States was provided through installment loans — fixed-term loans, extended by banks or by retailers, for the purchase of

specific goods. A consumer who wished to buy a refrigerator would apply for an installment loan at a bank or at the appliance store, would be approved or denied based on the bank's assessment of the consumer's creditworthiness, and would, if approved, receive the loan and repay it in fixed monthly installments over a period of one to three years. The system worked, in the sense that it provided credit. The system was, however, cumbersome — each loan required a separate application, a separate approval, and a separate set of paperwork. The BankAmericard, which was a revolving credit facility — a line of credit that could be used, repaid, and reused, without a new application for each purchase — was simpler, more convenient, and more flexible than the installment loan system it replaced. The convenience was the card's principal attraction. The convenience was also the source of its profitability — the revolving balances generated interest income, the merchant fees generated fee income, and the combination of the two produced, over time, a business that was substantially more profitable than the installment loan business it had replaced.

The BankAmericard, over the following decade, evolved into the Visa network. In 1970, the Bank of America, recognizing that the card's growth required a broader ownership structure than a single bank could provide, spun the card off into a cooperative — the National BankAmericard Corporation, later renamed Visa International — owned by the member banks that issued the cards and acquired the merchants. The cooperative structure was a recognition that the credit card business required network effects that no single bank could achieve alone. A card that was accepted by merchants in only one state was of limited use to a consumer who traveled. A card that was accepted by merchants in all states required a network of member banks, each issuing cards and acquiring merchants in its own territory, and each participating in a common settlement system. The cooperative provided the network. The member banks provided the issuance and the acquisition. The merchants provided the acceptance. The consumers provided the usage. The network effects — each new merchant making the card more valuable to consumers, each new consumer making the card more valuable to merchants — drove the system's growth. By the 1980s, Visa was the dominant credit card network in the United States, and its principal competitor, Master Charge (later Mastercard), which had been formed in 1966 by a group of competing banks, had adopted the same cooperative structure and was growing in parallel. The two networks — Visa and Mastercard — were, by the end of the 1980s, a duopoly. They have remained a duopoly ever since.

The competitive dynamics of the duopoly are not the dynamics that a textbook on industrial organization would predict. A duopoly, in the textbook model, competes — on price, on quality, on innovation — because each firm has an incentive to take market share from the other. The Visa-Mastercard duopoly does not compete on price. The interchange fees that the two networks set are, for any given card type and merchant category, substantially identical. Visa's interchange schedule and Mastercard's interchange schedule are, to the merchant who pays them, indistinguishable. The networks do not compete on price because the price is not a competitive variable. The price is set, in effect, by the industry — by the consensus of the networks and the issuing banks, who share an interest in maintaining the interchange at a level that generates substantial revenue for both. The competition between Visa and Mastercard is, instead, a competition for issuance — a competition to persuade the issuing banks to issue Visa-branded cards rather than Mastercard-branded cards, or vice versa. The

competition is conducted through the networks' relationships with the banks, through the incentives the networks offer the banks for issuance, and through the marketing campaigns the networks run to build brand awareness among consumers. The competition does not reduce the interchange fee. It distributes the interchange revenue between the two networks, in proportions that have remained roughly stable — Visa at roughly sixty percent, Mastercard at roughly forty percent — for the past three decades.

The duopoly's competitive stability is remarkable. The American payments market has, over the past five decades, seen the entry of numerous competitors — Discover, American Express, various store cards, various private label cards, and, more recently, a range of digital wallet providers and technology companies that have attempted to insert themselves into the payments flow. None of these competitors has materially eroded the duopoly's share of the general-purpose credit card market. American Express, which operates a closed-loop network (it is both the issuer and the acquirer of its cards, rather than relying on separate issuing and acquiring banks), has maintained a stable but minority share — roughly five to ten percent of the general-purpose card market — for decades. Discover, which also operates a closed-loop network, has a similar share. The digital wallets — Apple Pay, Google Pay, Samsung Pay — which allow consumers to make payments using their mobile phones rather than their physical cards, do not, despite their consumer-facing prominence, represent a competitive challenge to the duopoly. The digital wallets process transactions over the Visa and Mastercard networks, using cards issued by Visa and Mastercard issuing banks. The wallets are an additional layer on top of the duopoly, not a substitute for it. The duopoly's position, at the center of the American payments system, has been unchanged by the technological innovations of the past two decades. The position is durable — the investors who value the duopoly's shares at nearly a trillion dollars combined are betting it will hold.

* * *

The economics of the credit card business, as they evolved over the decades after the Fresno Drop, are the economics that the rest of this chapter is concerned with, and because they are, to most consumers, invisible. The invisibility is not accidental. The invisibility is the principal achievement of the industry's design.

When a consumer uses a credit card to make a purchase, the transaction involves four parties. The consumer is the first party. The merchant is the second party. The bank that issued the credit card to the consumer — the issuing bank — is the third party. The bank that acquired the merchant — the acquiring bank, which processes the card payments on behalf of the merchant and deposits the funds in the merchant's account — is the fourth party. The card network — Visa or Mastercard — sits in the middle, connecting the issuing bank to the acquiring bank, and setting the rules and the fees that govern the transaction.

The fee that sits at the center of this arrangement is the interchange fee. The interchange fee is paid by the acquiring bank to the issuing bank, for each transaction processed over the network. The fee is a percentage of the transaction amount — roughly two percent, on average, for credit card transactions in the United States — plus a fixed per-transaction charge of roughly ten cents. The fee is set by the card network — Visa or Mastercard — and is published, in a schedule that runs to hundreds of pages, in the network's operating regulations. The fee varies by card type (premium rewards cards carry higher interchange than standard cards), by merchant category (gas stations and grocery stores pay different rates than restaurants and airlines), and by transaction size (the percentage component is capped for very large transactions). The complexity of the schedule is considerable. The aggregate effect of the schedule, however, is straightforward: the issuing bank receives, from the acquiring bank, roughly two percent of every credit card transaction that its cardholders make.

The acquiring bank, which pays the interchange to the issuing bank, recovers the cost from the merchant. The merchant, which pays the acquiring bank a fee that includes the interchange plus the acquiring bank's own markup, recovers the cost from the consumer. The consumer, who pays the merchant the listed price of the goods, does not see the interchange fee. The fee is not itemized on the receipt. The fee is not disclosed at the point of sale. The fee is, from the consumer's perspective, invisible — embedded in the price of the goods, indistinguishable from the cost of the merchandise, the rent, the labor, and the other inputs that the merchant has priced into the product. The consumer pays the fee, without knowing that they are paying it, every time they use a credit card. The consumer also pays the fee, without knowing that they are paying it, every time they pay with cash, because the merchant, who must cover the average cost of accepting cards across all transactions, prices the interchange into the goods rather than surcharging it to card users specifically. The cash payer subsidizes the card payer. The card payer, if they use a rewards card, receives a portion of the interchange back in the form of cash back or travel points. The net effect is a transfer from cash payers (who tend to be lower-income) to rewards card holders (who tend to be higher-income). The transfer is substantial — tens of billions of dollars annually, in the United States alone, redistributed from lower-income consumers to higher-income consumers through a mechanism that neither group fully understands.

The invisibility of the interchange fee is the system's defining feature. The fee is, in economic substance, a tax on consumption, levied by the banks and the networks on every transaction that passes through the card system. The tax is the opposite of a sales tax — which is itemized on the receipt, visible to the consumer at the point of sale, and subject to political debate whenever a legislature proposes to raise it. The interchange tax is invisible, unitemized, and exempt from political debate, because the consumers who pay it do not know they are paying it, and the legislators who might regulate it hear, primarily, from the banks and the networks that collect it. The invisibility has been maintained, over five decades, through a combination of industry design (the fee is embedded in the price rather than surcharged at the point of sale) and industry lobbying (the industry has, whenever a legislature or a regulator has proposed to make the fee visible, opposed the proposal, on the grounds that visibility would confuse consumers and would disrupt the efficient functioning of the payments system).

The revenue that the interchange fee generates is one of the largest pools of consumer financial revenue in the American economy. The total interchange paid by American merchants in a typical year is estimated at roughly one hundred and sixty to two hundred billion dollars. This sum is divided, among the issuing banks (which receive the interchange), the acquiring banks (which receive their markup), and the card networks (which receive network assessment fees on the transaction volume), in proportions that vary by transaction type and by the relative bargaining power of the parties. The issuing banks receive the largest share — the interchange is a substantial source of revenue for the large card-issuing banks — JPMorgan Chase, Bank of America, Citibank, running into the tens of billions of dollars annually. The card networks — Visa and Mastercard — receive a smaller share, but their share is, because their operating costs are modest (they are, in essence, technology and branding companies that operate the network's switching infrastructure and set its rules), almost entirely profit. Visa's annual revenue, in a typical recent year, is roughly thirty billion dollars, of which a substantial fraction is net income. Mastercard's annual revenue is roughly twenty billion dollars, with a similar profit margin. The combined market capitalization of the two companies, as of the time of writing, is approximately nine hundred billion dollars — a valuation that reflects the market's assessment of the durability of the interchange fee system and the duopoly that operates it.

The duopoly has, over the five decades of its existence, defended the interchange system through a combination of mechanisms. The first mechanism is the network effect — the fact that any new entrant into the payments market must, to succeed, achieve acceptance among both merchants and consumers, and must do so against incumbents that already have universal acceptance. The network effect is, for Visa and Mastercard, a substantial barrier to entry. A new card network, however well-capitalized, cannot persuade a merchant to accept its card unless the merchant's customers want to use it, and cannot persuade consumers to carry its card unless merchants accept it. The chicken-and-egg problem is, for a new entrant, difficult to solve. The problem has, in fact, been solved only once in the past fifty years — by Discover, which was launched by Sears in 1985 and which achieved, over a period of decades, a market share of roughly five to ten percent, well behind Visa and Mastercard. No other entrant has achieved comparable scale.

The second mechanism is the relationship with the issuing banks. Visa and Mastercard are, despite their current status as publicly traded for-profit corporations, historically cooperatives — owned, until their initial public offerings in 2006 and 2008, by the member banks that issued the cards. The cooperative structure, which has been replaced by the publicly traded structure but which persists in the form of the banks' ongoing relationships with the networks, means that the banks that issue the cards are the banks that set the networks' policies. The banks' interest is in the preservation of the interchange fee, which is their revenue. The networks' interest is in the preservation of the banks' issuance, which is the networks' volume. The alignment of interests between the networks and the issuing banks has, over five decades, made the interchange system effectively unchallengeable from within the industry. The merchants, who pay the fee, have no seat at the table where the fee is set. The consumers, who pay the fee in the form of higher prices, do not know that they are paying it. The only parties that could challenge the fee — the legislators and the regulators — hear, when they consider challenging it, from the banks and the networks, whose resources and lobbying capacity substantially exceed those of the

merchants and the consumer advocates.

The third mechanism is the political one. Visa and Mastercard, together with the large issuing banks, maintain, in Washington, one of the most effective lobbying operations in the financial services industry. The lobbying is directed at the preservation of the interchange system — at preventing the passage of legislation that would cap interchange fees, at preventing the implementation of regulations that would make the fees visible, and at maintaining the regulatory framework that allows the networks to set the fees without supervision. The lobbying has been extraordinarily successful. The United States, alone among the major developed economies, has no cap on credit card interchange fees. The European Union caps them at 0.3 percent. Australia caps them at 0.5 percent. The United States caps only debit card interchange (through the Durbin Amendment), and the cap applies only to banks with more than ten billion dollars in assets, leaving the smaller banks — which account for a substantial fraction of debit issuance — exempt. The absence of a credit interchange cap in the United States is an anomaly. The anomaly is a consequence of the lobbying.

* * *

The Durbin Amendment, which was passed in July 2010 as part of the Dodd-Frank Wall Street Reform and Consumer Protection Act, was the most serious attempt, in the fifty-year history of the American credit card industry, to regulate the interchange fee. The amendment, which was introduced by Senator Richard Durbin, a Democrat from Illinois, applied to debit card transactions — not to credit card transactions — and required the Federal Reserve to set a "reasonable and proportional" interchange fee for debit transactions processed by banks with more than ten billion dollars in assets. The Federal Reserve, in implementing the amendment, set the fee at roughly twenty-one cents plus five basis points (0.05 percent) of the transaction value, which was a substantial reduction from the pre-Durbin average of roughly forty-four cents per transaction. The reduction, in aggregate, represented a transfer of several billion dollars annually from the issuing banks to the merchants.

The Durbin Amendment was not the first attempt to challenge the interchange system through legal or regulatory means. The merchants, who had chafed under the interchange fees for decades, had, in the early 2000s, brought a class action lawsuit against Visa and Mastercard, alleging that the networks' setting of interchange fees — through a process that involved the networks and their member banks, but not the merchants — constituted a price-fixing arrangement in violation of the antitrust laws. The lawsuit, which was led by Wal-Mart and a class of other large merchants, was settled in 2003 for roughly three billion dollars, plus an agreement by the networks to modify certain of their rules regarding debit card acceptance. The settlement was substantial. It was also negligible in structural effect — the interchange fees continued to be set by the networks, the merchants continued to pay them, and the system continued to operate as it had before. The lawsuit had, if anything, confirmed the system's resilience — the merchants had spent years and substantial legal fees challenging the

interchange, had obtained a large monetary settlement, and had not, in the process, changed the system's fundamental economics. The interchange system had survived the legal challenge, as it would survive the regulatory challenge that followed.

The Durbin Amendment's proponents, having observed the limited effects of the Wal-Mart settlement, pursued a regulatory rather than a legal route. The amendment, attached to the Dodd-Frank bill in the final days of its legislative drafting, was passed with minimal debate — the broader Dodd-Frank legislation, which was responding to the 2008 financial crisis, was considered under a procedural fast-track that limited amendments, and the Durbin Amendment was included with relatively little public discussion. The banking industry, which had not anticipated the amendment's inclusion, was caught off guard. The industry's lobbyists, who had been focused on the broader Dodd-Frank provisions, had not devoted substantial resources to opposing the interchange provision. The amendment passed, and the industry, having lost the legislative battle, shifted its resources to the regulatory battle — the Federal Reserve's implementation of the amendment's requirements.

The Federal Reserve's implementation disappointed the merchants. The amendment had required the Federal Reserve to set a "reasonable and proportional" fee, and the merchants had argued that a reasonable and proportional fee would be roughly twelve cents per transaction — the Federal Reserve's own estimate of the actual cost of processing a debit transaction. The Federal Reserve, in its final rule, set the fee at twenty-one cents plus five basis points, which was substantially higher than the merchants' preferred level. The Federal Reserve's justification was a recognition of the need to allow the issuing banks to recover not only the direct processing costs but also the indirect costs of fraud prevention, network operation, and other functions that the debit system required. The merchants saw it as a concession to the banks' lobbying. The Federal Reserve saw it as a balancing of competing interests — a recognition that the fee could not be set at the pure processing cost without jeopardizing the viability of the debit card system, which the Federal Reserve did not wish to do.

The Durbin Amendment was, by the stated objectives of its proponents, a measure designed to reduce the fees that merchants paid for debit transactions, in the expectation that the reduction would be passed through to consumers in the form of lower prices. The amendment's proponents argued that the interchange fee, which was set by the card networks without negotiation between the merchants and the banks, was a market failure — a price set above the competitive level by a duopoly that faced no effective competition. The regulation, the proponents argued, would correct the market failure, would reduce the merchants' costs, and would, through competitive pressure, result in lower prices for consumers.

The amendment's opponents — the card networks, the issuing banks, and their allies in Congress — argued that the regulation would not result in lower prices for consumers, because the merchants, having received the reduction in interchange, would not pass it through. The opponents also argued that the regulation would result in higher banking fees for consumers, because the banks, having lost the interchange revenue, would seek to replace it by charging depositors more for checking accounts, overdrafts, and other banking services. The opponents' prediction was accurate. The banks, in the years following the Durbin Amendment's implementation, did, in fact, raise banking fees. The proportion of banks offering free checking accounts with no minimum balance requirement declined, from roughly

three-quarters of banks before Durbin to roughly half after. The minimum balance requirements for free checking increased. The overdraft fees, which had already been a substantial source of revenue for the banks, were maintained or increased. The net effect of the amendment, as assessed by the economists who have studied it, was approximately zero for consumers — the savings from lower debit interchange were offset by the increased banking fees, with the result that consumers, in aggregate, paid roughly the same amount as they had paid before the amendment, but paid it in different forms and to different recipients.

The Durbin Amendment's failure to produce net benefits for consumers has been a cautionary tale for the reformers who subsequently attempted to extend interchange regulation to credit cards. The Credit Card Competition Act, which was introduced in 2022 and reintroduced in 2023, would have required the large issuing banks to enable at least two competing networks on each credit card they issue — one of which could not be Visa or Mastercard. The requirement was designed to introduce competition into a market that, under the current duopoly structure, has none. The Act was supported by a coalition of merchants, merchant associations, and consumer advocates. The Act was opposed by the card networks, the issuing banks, and their allies in Congress. The Act did not pass. The Act has not, as of the time of writing, passed, and the prospects for its passage are limited.

The failure of the Credit Card Competition Act, and the broader failure of every attempt to regulate credit card interchange in the United States, is a pattern of institutional resilience. The Visa-Mastercard duopoly has, over five decades, built an interchange system that extracts, from the American economy, roughly two hundred billion dollars annually. The system is a tax — a levy on consumption, borne by all consumers, collected by the banks and the networks. The system is effectively unreformable. The reformers have, for decades, attempted to reduce or to eliminate the tax. The reformers have failed. The failure is not a consequence of the reformers' incompetence, or of the merits of their proposals. The failure is a consequence of the institutional position of the duopoly — the fact that the duopoly's revenue is embedded in the economics of every American merchant and every American bank, and that any attempt to reduce the revenue produces opposition from the parties that receive it.

The contrast with the European Union, which capped interchange at 0.3 percent for credit and 0.2 percent for debit in 2015, is instructive. The European cap was implemented through regulation, by the European Commission, over the opposition of the card networks and the European banks. The cap has been in effect for nearly a decade. The European payments system continues to function. The European banks continue to issue cards. The European merchants continue to accept them. The European consumers continue to use them. The reduction in interchange has not, contrary to the predictions of the cap's opponents, destroyed the European card market. The reduction has, however, reduced the revenue of the European banks and the European operations of the card networks, which was the cap's purpose. The cap was politically feasible in Europe because the European banks, operating in a regulatory environment that is less sympathetic to bank profitability than the American environment, were unable to prevent it. The cap has not been politically feasible in the United States, because the American banks, operating in a regulatory environment that is more sympathetic to bank profitability, have been able to prevent it.

The reasons for the political divergence illuminate the way that institutional architecture, rather than policy merit, determines outcomes. The European Commission, which proposed and implemented the interchange cap, is a regulatory body that operates, in the domain of competition policy, with a degree of independence from the industries it regulates. The Commission's competition directorate has, over the past two decades, taken a series of actions against American technology companies, European banks, and other large enterprises, on the grounds that their practices violated European competition law. The actions have been, in their aggregate effect, a assertion of European regulatory authority over practices that, in the American context, would be regarded as within the normal operation of the market. The European regulatory architecture, in other words, is more willing to intervene in the pricing decisions of private firms than the American regulatory architecture is. The willingness is a product of the European political tradition, which regards the regulation of large enterprises as a legitimate and necessary function of the state, in a way that the American political tradition, which regards the regulation of private enterprise with more suspicion, does not.

The American regulatory architecture, in the domain of interchange, is characterized by a different set of assumptions. The Federal Reserve, which has the authority to regulate certain aspects of the payments system, has historically been reluctant to intervene in the pricing of interchange, on the grounds that the pricing is a matter for the market to determine. The Federal Reserve's intervention, through the Durbin Amendment, was limited to debit interchange, and was undertaken at the direction of Congress rather than on the Federal Reserve's own initiative. The Federal Reserve has not, on its own initiative, extended the regulation to credit interchange. Congress, which could direct the Federal Reserve to do so, has not done so, because the members of Congress who would sponsor such legislation face, in the card networks and the issuing banks, opponents with substantial lobbying resources and with a demonstrated capacity to defeat or to water down legislation that threatens their revenue. The Credit Card Competition Act, which was the most recent attempt, failed to advance in both 2022 and 2023, despite the support of a coalition of merchants that included some of the largest retailers in the United States. The failure was not a consequence of the Act's merits, which were, by the economists' assessment, substantial. The failure was a consequence of the political arithmetic of the United States Congress, in which the banks and the networks have, through campaign contributions, lobbying expenditures, and the cultivation of relationships with key legislators, maintained a position that the merchants and the consumer advocates have been unable to overcome.

The difference between the European and the American outcomes is not, primarily, a difference in the economics of the payments system. The economics are, in both cases, the same — a duopoly of card networks, setting interchange fees, collecting revenue from merchants, and passing it to issuing banks. The difference is in the political economy — the relative power of the banks and the networks, relative to the power of the merchants and the consumers, in the legislative and regulatory processes that determine the fees. In Europe, the merchants and the consumers prevailed. In the United States, the banks and the networks prevailed. The outcome, in each case, was determined not by the merits of the policy but by the institutional architecture of the political system that produced it.

The Chase Sapphire Reserve, launched in 2017 by JPMorgan Chase, illustrates it illustrates the way the interchange system, far from being a static relic of the 1960s, continues to evolve and to generate

new revenue streams. The Sapphire Reserve was, at its launch, a premium travel rewards card, with a four hundred and fifty dollar annual fee (later raised to five hundred and fifty, then six hundred and ninety-five), a one hundred thousand point sign-up bonus (worth roughly fifteen hundred dollars in travel), and a rewards structure that offered three points per dollar spent on travel and dining. The card was targeted at affluent millennials — the demographic that was beginning to earn enough to qualify for premium credit products but had been slow to adopt them. The marketing worked. The response to the launch exceeded Chase's projections so dramatically that the bank, in the weeks following the launch, ran out of the metal stock used to manufacture the cards. The shortage — which became, in the financial press, a symbol of the card's popularity — was the most honest moment in the interchange story. The demand for the card was a demand for a device that would transfer interchange revenue, extracted from merchants, to the cardholder in the form of rewards. The cardholders who applied were not aware of the interchange mechanics that funded the rewards. They were aware of the rewards. The rewards were funded by the merchants. The merchants passed the cost to all consumers. The consumers, in aggregate, paid for the rewards that the Sapphire Reserve cardholders received. The transfer was invisible. It was also substantial.

The Sapphire Reserve was not an isolated phenomenon. It was one of a class of premium rewards cards — including the American Express Platinum, the Citi Prestige, and various other products — that have, over the past decade, proliferated in the American market. The cards offer, in their aggregate, rewards that are funded by interchange fees that are higher than the interchange fees on standard cards. The premium cards carry higher interchange rates — the merchants who accept them pay more per transaction than they pay for standard cards — and the additional revenue is passed, through the rewards structure, to the cardholders. The premium card business is a mechanism for extracting additional interchange from merchants and distributing it to affluent cardholders. The mechanism is uncontroversial — the premium cards are regarded as desirable products, and the cardholders who carry them are regarded as savvy consumers who have taken advantage of the rewards system. The merchants, who pay the higher interchange, have no effective means of refusing the cards — the network rules require merchants that accept Visa or Mastercard to accept all Visa or Mastercard cards, including the premium cards with the higher interchange. The merchants' only option, if they wish to avoid the higher interchange, is to stop accepting the network's cards entirely, which would, for most merchants, result in a loss of business that exceeds the cost of the interchange. The merchants accept the cards. The interchange is paid. The rewards are distributed. The system continues.

The system's continuation is the point. The Visa-Mastercard duopoly is, unlike the other institutions described in this book, an institution that has not failed. It has not been outpaced by technology, or bypassed by a competitor, or rendered obsolete by a change in the market. It has adapted to every change, defended its position against every challenge, and continued, year after year, to extract its two percent from the American economy. The extraction is a success story — an institution that has matched the velocity of the market, responded to every threat with an effective defense, and preserved its rent extraction through five decades of technological and political change. The success is a triumph for the duopoly's shareholders. It is a cost for the consumers and merchants who pay the two percent — invisible to most of the people who pay it, and demonstrated, by the rest of the world, to be

unnecessary. The Indian UPI, the European interchange cap, the Brazilian Pix have all shown that a payments system can function without the two percent. The demonstration has not, so far, produced a reform in the United States. It may, in time. The time, on the evidence of the past five decades, will be considerable.

The Fresno Drop, in retrospect, was the beginning of a system that now processes trillions of dollars in transactions annually, that generates hundreds of billions of dollars in fees, and that has, over six decades, become so embedded in the American economy that its reform is effectively impossible. The sixty thousand cards that the Bank of America mailed to the residents of Fresno in 1958 have, over the following decades, multiplied into the billions of cards that are in circulation today, and the network that was bootstrapped by that initial mailing has become the duopoly that sits, invisibly, at the center of every American consumer's daily transactions. The duopoly is the most successful institution described in this book. It is also the most expensive. The two facts are not unrelated. The success has been built on the expense, and the expense has been made invisible — maintained through fifty years of design, lobbying, and the quiet accumulation of the most effective political position in the American financial system. The maintenance continues.

Chapter 11

The Coldplay Dollar

In the year 2022, a person in Buenos Aires who wished to exchange pesos for dollars could transact at any of the following exchange rates: the official rate, published daily by the Central Bank, which everyone who read it knew to be a fiction; the blue rate, quoted on a website called dólarblue.net.ar, which everyone who used it knew to be the actual price; the MEP rate, derived from the price of Argentine government bonds denominated in dollars and traded on the Buenos Aires stock exchange; the crypto rate, quoted on cryptocurrency exchanges and roughly equal to the blue rate; the soybean rate, offered to soybean exporters to encourage them to repatriate their dollar earnings; the tourist rate, the official rate plus a tax surcharge applied to Argentines who used their credit cards abroad; the Qatar rate, a more punitive surcharge applied to Argentines who traveled to the World Cup; and the Coldplay rate, a special exchange rate applied to Argentines who purchased tickets to the Coldplay concerts held in Buenos Aires in October 2022.

These are not inventions. Each was, in 2022, an official exchange rate, published by the central bank or by a financial authority operating under its supervision, and applied to a specific category of transaction. The rates were the product of a government's attempt to manage a currency it could not manage — by inventing a separate price for each use to which the currency was put. They formed a labyrinth of corridors, each leading to another corridor, each labeled with a rate that was real in the sense that it was quoted and transacted, and fictional in the sense that it did not describe a single price for the peso.

An Argentine who wished to buy dollars for savings went to a cueva — an informal exchange house, typically operating out of the back of a clothing store — and transacted at the blue rate. An Argentine who wished to buy dollar-denominated bonds bought them through a brokerage account and transacted at the MEP rate. An Argentine who traveled abroad used a credit card and was charged at the tourist rate, or, if traveling to Qatar for the World Cup, at the Qatar rate. An Argentine who attended a Coldplay concert bought a ticket and was charged at the Coldplay rate. An Argentine who exported soybeans repatriated the proceeds at the soybean rate. Each price was real. Each price was incompatible with the others. Each was maintained by a central bank that had, over the preceding two decades, lost the ability to maintain a single price for its own currency.

The Coldplay rate is the rate that most clearly reveals the system. Coldplay is a British rock band, which in 2022 embarked on a world tour that included ten concerts at the River Plate stadium in Buenos Aires. The ten shows sold out. The demand for tickets was substantial. The tickets were priced in dollars, because the concert promoters were an international entity that wished to be paid in a currency it trusted. The Argentine fans who wished to buy the tickets had pesos, because they were paid in pesos, because they lived and worked in Argentina. The conversion of pesos to dollars, under the official exchange rate, would have provided the fans with a subsidy — the official rate was more favorable than the blue rate, and a fan who converted at the official rate would have paid, in dollar terms, substantially

less than a fan who converted at the blue rate. The subsidy was a cost to the government — it was, in effect, selling dollars to concert-goers at a price below the market price, and the difference was borne by the central bank's reserves. The government, unwilling to bear the cost of allowing Argentines to attend a British rock concert, invented a new exchange rate. The Coldplay rate was set between the official rate and the blue rate — less favorable than the official rate, more favorable than the blue rate — and applied exclusively to the purchase of Coldplay tickets. The rate was announced, published, and used. Tickets were purchased. The transactions cleared. The Coldplay dollar existed, for a period of weeks, as a currency — a specific price for the peso, maintained by the central bank, to regulate the flow of pesos into the bank accounts of a British rock band.

The Coldplay dollar is not the absurdity of a government that has lost its mind. It is the absurdity of a government that has lost its currency. The Argentine peso, in 2022, did not have a single price, because the central bank could not maintain a single price. The central bank could not maintain a single price because the Argentine public did not believe the price the central bank published, and because the public's disbelief was self-fulfilling — the refusal to hold pesos, and the preference for dollars, produced the inflation and the devaluation that made the published price fictional. The central bank, unable to maintain a single price, attempted to maintain multiple prices — one for each category of transaction it wished to influence. The multiple prices were a confession that the central bank had lost control, dressed up as a regulatory framework. The confession was familiar. The Argentine public had been listening to it, in one form or another, for twenty years.

The labyrinth of rates was a thing that no novelist would have dared invent, because the invention would have been dismissed as implausible. A government that maintains seven exchange rates simultaneously, each applied to a different category of transaction, each published daily as if it were the price, and each different from the others — this is not a government that is managing its currency. This is a government that is administering its decomposition. The officials publish the rates. The rates are quoted. The transactions are processed. The system functions, in the narrow sense that the transactions clear. It does not function, in the broader sense that the currency it administers has ceased to be a store of value. It functions and does not function, simultaneously — the narrative proceeds, the details accumulate, and the ground beneath it shifts, gradually, until you realize that the story has been describing, all along, a world that does not exist.

The Coldplay dollar was a perfect artifact of this world. The central bank, faced with the prospect of Argentines spending pesos to attend a rock concert, invented a currency. It was not, technically, a new peso. It was a rate — a specific price at which pesos would be converted to dollars for the specific purpose of buying Coldplay tickets. The rate was real, in the sense that it was quoted and transacted. It was fictional, in the sense that it did not describe the value of the peso. It was a Tlön — a thing imagined into being by an act of declaration, sustained for a period of weeks by the collective agreement of the people who transacted in it.

* * *

The story of the Argentine peso's decline, from a currency that was, for a decade, as good as the dollar, to a currency that is, in 2024, held by no one who can avoid holding it, is a story about trust. Trust is the commodity that money is made of. A currency is valuable because the people who hold it believe that it will be valuable tomorrow, and because the belief, held by enough people, makes the currency valuable today. The belief is circular — the currency is valuable because people believe it is valuable, and people believe it is valuable because the currency is valuable. The circularity is invisible when a currency works. It is the thing everyone talks about when a currency has stopped working.

The Argentine peso stopped working, in the modern iteration of its decline, in December 2001. The peso had, since 1991, been pegged to the dollar at a rate of one to one, under a system called the Convertibility Plan. The plan, implemented by the government of Carlos Menem in response to the hyperinflation of 1989 and 1990, was a renunciation of monetary policy. The Central Bank of Argentina, under the Convertibility Plan, was required to hold dollar reserves equal to its peso liabilities, and was prohibited from issuing pesos except in exchange for dollars at the fixed rate. The plan was, in effect, a dollarization — the peso became a dollar derivative, and the Argentine economy operated, for a decade, on a currency that was, for all practical purposes, the dollar.

The Convertibility Plan worked, for a decade, because it was credible. The Argentine public, which had been burned by hyperinflation in 1989, accepted the peg as a commitment — a promise by the government that the peso would be worth a dollar, today and tomorrow and forever. The promise was kept, for a decade. The peso was worth a dollar. Inflation, which had been 5,000 percent in 1989, fell to single digits by 1993, and remained there for the rest of the decade. The Argentine economy grew. Foreign investment flowed in. Argentine banks, which had been hollowed out by the hyperinflation, recapitalized. Argentine companies, which had been unable to borrow in their own currency because no one trusted the currency, borrowed in dollars, at dollar interest rates, and used the proceeds to invest in Argentine projects. The decade of the 1990s was, in the Argentine memory, "los años dorados" — the golden years. The golden years were the product of a currency that people trusted.

The trust broke in December 2001. The break was a consequence of the Argentine government's inability to service its debt. The government, which had borrowed extensively in dollars during the 1990s, found, by 2001, that its dollar revenues (from exports and from taxes) were insufficient to cover its dollar obligations (to bondholders and to international creditors). The insolvency was, first, a fiscal problem — the government had spent more than it earned, and the gap had been bridged by borrowing that could not be repaid. The insolvency became, second, a monetary problem, because the Argentine public, which had been holding pesos on the assumption that the pesos were worth dollars, began to doubt the assumption. The doubt produced a run on the banks — depositors, fearing that the peg would break and that their pesos would lose their dollar value, attempted to withdraw their deposits in dollars. The government, unable to honor the withdrawals (because the central bank did not have enough dollars to pay out all the peso deposits at the one-to-one rate), froze the accounts. The freeze, known as the "corralito" — the little corral — was announced on December 1, 2001, and prohibited depositors from withdrawing more than a small weekly amount from their accounts. The corralito produced, in the

days that followed, the protests, the looting, the deaths, and the resignation — by helicopter, from the roof of the presidential palace — of President Fernando de la Rúa. The peg broke, formally, in January 2002. The peso was devalued, the exchange rate floated, and the peso, which had been worth one dollar for a decade, fell to four pesos per dollar within months.

The break of the peg was a trauma that has not healed. The public had trusted the peso, because the government had promised that the peso was worth a dollar. The government had broken the promise. The public had lost, in the corralito, access to its own savings, and had watched, in the devaluation, as the savings that remained lost three-quarters of their value. The lesson the public drew was simple, and it has governed Argentine monetary behavior for the two decades since: do not hold pesos. Hold dollars. Hold dollars in cash, under the mattress, in a safe deposit box, in a hidden compartment in a closet, in an account abroad, in any form that the Argentine government cannot reach, cannot freeze, cannot devalue. The lesson was learned, by every Argentine who could afford to learn it, in the years after 2002. The lesson has been passed, by those who learned it, to their children, who have grown up in a country where the assumption is not that the peso will hold its value, but that it will not.

The corralito deserves a moment of description, because it is the event that, more than any other, shaped the Argentine relationship with money. On December 1, 2001, the government of Fernando de la Rúa announced that bank accounts would be frozen — that depositors would be unable to withdraw more than a small weekly amount (initially 250 pesos, equivalent to 250 dollars under the peg) from their accounts. The freeze applied to all accounts, including dollar-denominated accounts, including accounts that the depositors had, over the preceding decade, regarded as their property. The legal justification was a measure to prevent a bank run — depositors had been withdrawing their deposits at a rate that the banks could not sustain, and the government, unable to provide the banks with the dollar liquidity they needed to honor the withdrawals, froze the accounts instead. The economics were sound. The politics were catastrophic. The depositors, who had been told by the government, for a decade, that their deposits were safe and that their pesos were worth dollars, discovered that the government's promises were, in a crisis, contingent. The promises were contingent because the government, which had guaranteed the deposits, did not have the dollars to pay them. The discovery produced, in the weeks that followed, the protests, the looting, the deaths — twenty-seven people died in the protests of December 2001 — and the resignation of President de la Rúa, who fled the presidential palace by helicopter on December 20, while the crowds below chanted for his departure.

The corralito was, in its aftermath, lifted — the accounts were unfrozen, the deposits were returned, but the deposits were returned in devalued pesos, at the rate of 1.4 pesos per dollar for amounts up to a threshold, and at the floating rate for amounts above. The depositors who had held dollar deposits received pesos, and the pesos they received were worth, within months, roughly twenty-five cents on the dollar. The loss — roughly seventy-five percent of the depositors' savings, for those who had held their deposits through the corralito — was a confiscation. The confiscation was a necessity — the government did not have the dollars to honor the deposits at their original value. The necessity was the lesson. The government's promises about money were not to be trusted. The lesson was learned, and the lesson has not been unlearned, in the two decades since.

The consequence of the lesson is that the Argentine economy operates, in substantial part, on dollars. The dollars are not, in the main, in the banking system — the banking system, after the corralito, is trusted by no one who remembers the corralito. The dollars are in cash, held by households, in quantities that the economists who study such things estimate at between two hundred and four hundred billion dollars. The sum is extraordinary. It exceeds the foreign exchange reserves of the Central Bank of Argentina. It exceeds the deposits in the Argentine banking system. It represents, on a per capita basis, one of the largest holdings of foreign currency cash in the world. The Argentine household, distrustful of the banking system and of the currency, has become its own central bank — holding its reserves, in cash, in the currency it trusts, in the place it trusts, which is home.

The central bank, faced with a public that holds dollars and that refuses to hold pesos, has spent two decades attempting to manage a currency that the public does not want. The management has taken the form of capital controls — the "cepo," or clamp — which restrict the ability of Argentines to purchase dollars through the formal financial system. The cepo, which was first imposed in 2011, limits each Argentine to the purchase of a small monthly dollar amount, at the official rate, through the banking system. The limit is an annoyance. The annoyance is circumvented, by the Argentine who can afford it, through the cueva — the informal exchange house, operating out of the back of a clothing store, which quotes the blue rate and which converts pesos to dollars without documentation, without limits, and without the central bank's knowledge.

The cueva deserves description, because it is the most remarkable feature of the Argentine monetary landscape. A cueva is, from the street, indistinguishable from any other small business. The sign on the door may say "boutique" or "agencia de viajes" or nothing at all. The front room is a clothing store, or a travel agency, or an office with a desk and a few chairs. The customer enters, asks for the manager, and is led — through a door, down a hallway, into a back room — to the operation. The back room contains a desk, a calculator, a computer displaying the current blue rate, and stacks of hundred-dollar bills. The customer states the amount he wishes to exchange. The cueva's operator quotes the rate. The customer hands over pesos. The cueva's operator counts out dollars. The transaction takes under five minutes. It is technically prohibited by the cepo. It is also universal. Every Argentine adult who can afford to hold dollars has, at some point, used a cueva. The cueva is what the ATM is to the American middle class — a routine mechanism for obtaining currency, relied upon without reflection, and discussed, when discussed at all, with the casual familiarity of a thing that has always been there.

The cueva's operator is a more reliable central banker than the central banker. The cueva's operator publishes, each morning, a rate that reflects the actual supply and demand for dollars in the Buenos Aires market. The rate is acknowledged by every participant in the market — by the households that check it on their phones, by the businesses that use it to price their contracts, by the economists who cite it in their analyses, and, not least, by the central bank itself, whose officials consult the blue rate, in private, to assess the gap between their published fiction and the market's reality. The cueva's operator does not set the rate by fiat. The rate is set by the market — by the supply of pesos seeking dollars and the supply of dollars seeking pesos, in the informal market that the cuevas collectively constitute. The operator is a price-taker, not a price-maker. The price he quotes is the price the market has arrived at,

through the aggregation of thousands of individual transactions, each conducted in a back room, each invisible to the central bank, and each contributing, in its small way, to the discovery of the actual price of the peso. The central bank, which publishes a different price, is, by comparison, a price-pretender — an institution that publishes a price that the market does not accept, and that maintains the publication, day after day, as if the publication could make the price real. The publication cannot make the price real. The price is made real by the transactions, and the transactions are conducted in the cuevas, at the blue rate, by the public that has, for two decades, given up on the central bank's ability to set a price that bears any relation to reality.

* * *

The proliferation of exchange rates — the labyrinth that the Argentine monetary system became, in the years after 2012 — was the central bank's response to a problem it could not solve. The problem was that the peso had ceased to be a store of value. The public held pesos only for the brief period necessary to spend them, and converted pesos to dollars at the first opportunity. The conversion, which the central bank wished to discourage (because each conversion depleted the central bank's dollar reserves), was restricted by the cepo. The cepo limited the formal conversion, but could not limit the informal conversion — the cueva, the blue rate, and the parallel market that operated outside the central bank's control. The central bank, unable to prevent the conversion, attempted to manage it, by inventing a separate exchange rate for each category of conversion it wished to influence.

The soybean rate was invented in 2012, to encourage soybean exporters to repatriate their dollar earnings. The exporters, who sold soybeans on the international market for dollars, had been holding their dollars abroad, rather than bringing them into Argentina, because bringing dollars into Argentina meant converting them to pesos at the official rate, which was unfavorable. The central bank, which needed the dollars to replenish its reserves, offered the exporters a special, more favorable rate — the soybean rate — as an incentive to repatriate. The rate worked, in the sense that some dollars were repatriated. The rate also worked, in the sense that it confirmed to the Argentine public that the official rate was a fiction — if the central bank was willing to offer a different rate to soybean exporters, then the official rate was not the price of the peso, but one price among several, each applied to a different category of transaction.

The tourist rate was invented to address a different problem. Argentines who traveled abroad used their credit cards, which were charged at the official rate. The official rate, being more favorable than the blue rate, represented a subsidy — the Argentine who used a credit card abroad was, in effect, buying dollars at a price below the market price, and the difference was borne by the central bank's reserves. The central bank, unwilling to subsidize foreign travel, added a tax surcharge to credit card transactions abroad, producing the tourist rate — a rate that was less favorable than the official rate, and that was a recognition that the official rate could not be maintained for transactions that the central bank

found politically inconvenient.

The Qatar rate was a further refinement. The World Cup, which was held in Qatar in November and December 2022, produced a surge of Argentine travel — Argentines, who are enthusiastic about football, traveled to Qatar in large numbers to support their national team. The travel, which involved credit card spending abroad, would have been subsidized at the official rate. The central bank, unwilling to subsidize World Cup travel, invented the Qatar rate — a rate that was more punitive than the tourist rate, applied specifically to Argentines traveling to Qatar. The rate was announced, published, and applied. The Argentines who traveled to Qatar paid at the Qatar rate. The rate existed, for the duration of the World Cup, as a currency — a specific price for the peso, maintained by the central bank, to regulate the flow of pesos into the Qatari economy.

The Coldplay rate, which was announced in October 2022, was the most specific of the rates, and the most absurd. The rate was applied exclusively to the purchase of tickets to the Coldplay concerts in Buenos Aires. The concerts, which were ten in number, were held at the River Plate stadium, which has a capacity of roughly sixty thousand. The tickets were sold through a website, which accepted payment in pesos, at the Coldplay rate. The rate was between the official rate and the blue rate — less favorable than the official rate, more favorable than the blue rate. The level was chosen to balance two considerations: the concert promoters, who wished to be paid in dollars at the market rate, and the Argentine fans, who wished to pay in pesos at a rate they could afford. The Coldplay rate was the compromise. The compromise was a monetary policy — a decision by the central bank to set a specific price for the peso, to regulate the purchase of tickets to a British rock concert. The policy was announced, implemented, and concluded. The concerts were held. The tickets were purchased. The Coldplay dollar ceased to exist, as a rate, when the concerts ended. The Coldplay dollar had been as real as any other currency — quoted, transacted, and cleared through the financial system. The Coldplay dollar had also been as fictional as any other currency — a price set by a central bank for a purpose that had nothing to do with the value of the peso, and everything to do with the central bank's inability to set a single price for the peso that the public would accept.

The proliferation of rates was a solution to a problem that the central bank could not solve. The problem was that the peso had no single price, because the public did not believe in the peso. The solution was to invent multiple prices, each for a specific purpose, each maintained as if it were the price. The solution produced a multiplication of fictions — a labyrinth of rates, each corridor labeled with a price, each price real in the sense that it was transacted, and fictional in the sense that it did not describe a single value for the currency. The Argentine who navigated the labyrinth was a visitor to a library of prices, each price a book in the library, each book describing a different world, and each world existing, simultaneously, in the same city, in the same economy, in the same currency.

The labyrinth could not, by its nature, endure. The multiplication of rates was, at best, a temporary management of a problem that the central bank could not solve. The problem was that the peso was not trusted, and no amount of rate multiplication could restore the trust. The trust was gone, and had been gone since 2001, and the Argentine public had, in the two decades since, built a parallel monetary system — the cueva, the blue rate, the dollar under the mattress — that operated without the central bank and without the peso. The parallel system was larger than the formal system. The dollars held by

Argentine households, in cash, exceeded the reserves of the central bank. The transactions conducted at the blue rate exceeded, by some estimates, the transactions conducted at the official rate. The parallel system was the monetary system of Argentina, and the formal system was a fiction that the central bank maintained for the benefit of the international organizations that lent Argentina money and that required, as a condition of the lending, the existence of a central bank and an official exchange rate.

The election of Javier Milei, in November 2023, was a recognition that the fiction could no longer be maintained. Milei, who is a libertarian economist and a former television commentator, campaigned on a platform that included the abolition of the central bank and the adoption of the United States dollar as Argentina's currency. The platform was an acknowledgment that the Argentine peso had failed, and that the failure was so complete that the currency should be abandoned. The platform was successful — Milei won the election, with roughly fifty-six percent of the vote, in a runoff against Sergio Massa, the minister of economy. The victory was a signal that the era of the peso was ending, and that the central bank's days were numbered.

The crypto rate — the dólar cripto — appeared in the catalog of Argentine exchange rates in the late 2010s, and it illustrates the way that the Argentine public, having lost faith in the peso and in the banking system, has adopted whatever technology offers an alternative. The crypto rate is the price of a dollar-pegged stablecoin — typically USDT — on a cryptocurrency exchange, expressed in pesos. An Argentine who wishes to convert pesos to dollars can, instead of visiting a cueva, purchase USDT on a cryptocurrency exchange, paying in pesos, and hold the USDT in a digital wallet. The transaction settles in minutes, requires no physical visit to a back room, and leaves no record in the Argentine banking system. The crypto rate is approximately equal to the blue rate — the market, whether conducted in a cueva or on a blockchain, arrives at approximately the same price for the peso. The adoption of cryptocurrency in Argentina is, by several measures, among the highest in the world. The adoption is not driven by ideological enthusiasm for cryptocurrency. The adoption is driven by the same impulse that drives the cueva — the desire to hold dollars, in any form that the central bank cannot reach. The cryptocurrency, to the Argentine who uses it, is not a speculation. The cryptocurrency is a dollar. The cryptocurrency is the same thing that the hundred-dollar bill under the mattress is — a store of value, held outside the formal financial system, in a currency that the central bank cannot devalue.

The Argentine public, faced with a central bank that cannot maintain a currency, has adopted every alternative that technology has provided — the cueva, the blue rate, the dollar under the mattress, the cryptocurrency exchange. Each is mechanically different. Each does the same thing: holds value outside the central bank's reach. The alternatives have grown, alongside the central bank's decline, until they constitute a monetary system larger than the one the central bank nominally operates. The central bank publishes its rates. The public ignores them. The public uses the alternatives — which work.

Milei, upon taking office on December 10, 2023, did not immediately abolish the central bank or dollarize the economy. He did devalue the official exchange rate by roughly fifty-four percent — from three hundred and sixty-five pesos per dollar to eight hundred. It was the largest single-day adjustment in the peso's history. It brought the official rate closer to the blue rate, narrowing the gap. It did not eliminate the gap. The labyrinth persisted, reduced. The cepo persisted, modified. The central bank

persisted, diminished. The dollarization Milei had promised was, as of the time of writing, not yet implemented, and the peso, though devalued, remained the currency of Argentina.

The Argentine who holds dollars under the mattress has been proven right. The pesos she did not hold — converted to dollars at the blue rate, through the cueva, stored in cash — have retained their value. The pesos she did hold — kept in the banking system, at the official rate — have lost, over two decades, roughly ninety-nine percent of their value against the dollar. The lesson, learned in 2001 and relearned in every subsequent crisis, is that the peso is not money. It is a transactional convenience — accepted within Argentina, for the brief period between receipt and expenditure. The store of value is the dollar, held in cash, under the mattress, in the back of a closet, in a safe deposit box, in the place the central bank cannot reach.

Every morning, the Argentine who can afford to hold dollars checks the blue rate on her phone. The rate is published by no central bank. It is acknowledged by every participant in the economy. It is set, not by a government, but by the market that operates in the back rooms of clothing stores, on a website called dólarblue.net.ar, and in the collective judgment of millions of people who have, for two decades, been pricing the peso without the central bank's permission. The central bank publishes its own rate, daily, as if the rate were a fact. The public checks the blue rate, daily, as if the official rate were not.

Chapter 12

The Patchwork

On a Tuesday morning in the spring of 2025, a woman in Kibera sent two thousand shillings to her mother. The mother lived in a village three hundred kilometers to the west, near the shore of Lake Victoria. The woman was a domestic worker. She earned twenty thousand shillings a month, which is about a hundred and fifty American dollars. The two thousand she sent was a tenth of that.

She sent it by M-Pesa, on a Nokia phone she had bought used for two thousand shillings — the same amount she was now sending to her mother. She opened the application, selected "send money," entered her mother's phone number, entered the amount, and confirmed with a PIN. The two thousand shillings were debited from her account and credited to her mother's. The transaction settled before the woman had finished walking from her employer's house to the bus stop. The fee was thirty shillings, or about twenty cents. No bank was involved. The money moved on a system operated by Safaricom, a telecom, and was visible, in aggregate, to the Central Bank of Kenya — which regulated the system under a category that had not existed when it was launched.

The woman's mother was sixty-three years old. She had never had a bank account. Her mother — the domestic worker's grandmother — had also never had a bank account. The family had, for three generations, operated outside the formal financial system, not by choice but by exclusion. The banks had not wanted their business. The minimum balances were too high, the branches too far, the paperwork too complex, and the fees too large for the sums involved. The family had managed with cash — cash carried by bus, cash entrusted to neighbors, cash hidden in the house. The cash had been stolen, twice. Once from the grandmother's house, in a burglary that took everything she had saved for a year. Once from the mother, on a bus to Kisumu, where a pickpocket took the equivalent of two weeks' wages. The cash economy had been, for three generations, the only economy available, and the cash economy was dangerous.

The M-Pesa system had changed this. The domestic worker's mother had received her first M-Pesa transfer in 2009, from a daughter who was working in Nairobi. The transfer had been for five hundred shillings — about six dollars. The mother had walked to the M-Pesa agent, a woman who ran a small shop that sold airtime and sugar, and had collected the cash. The agent had counted out the notes, and the mother had folded them and placed them in her dress. The transaction had taken less time than the walk to the agent. The fee had been less than the cost of the bus fare to the nearest bank branch, which was forty kilometers away. The mother had used M-Pesa ever since. She had never opened a bank account. She had never needed to.

The two thousand shillings that the domestic worker sent on that Tuesday morning arrived at her mother's phone before the domestic worker had reached the bus stop. The mother was notified by a text message. She read the message, noted the amount, and set the phone down. She would collect the cash later, from the agent. The money would wait.

* * *

In Bangalore, two and a half hours ahead of Nairobi, a software engineer paid a street vendor ten rupees for a cup of tea. The vendor's cart was parked on a corner in Koramangala, near the offices where the engineer worked. The cart was wooden, with a glass front behind which the vendor displayed samosas, biscuits, and a large aluminum pot of tea brewed over a charcoal burner. Taped to the side of the cart was a piece of cardboard with a QR code printed on it. The engineer scanned the code, entered ten rupees — twelve American cents — and confirmed with a PIN. The money moved from his bank account to the vendor's bank account in a few seconds. The fee was zero. The transaction passed through the Unified Payments Interface, a rail operated by the National Payments Corporation of India, a non-profit consortium regulated by the Reserve Bank of India. It was one of roughly fourteen billion transactions that UPI would process that month — more than Visa and Mastercard combined.

The vendor did not know what the National Payments Corporation of India was. He did not know what a non-profit consortium was, or what the Reserve Bank of India had to do with the QR code on his cart. He knew that the code worked, that the money arrived, and that the system charged him nothing. He knew that the alternative — cash — required him to carry change, to guard against theft, and to make deposits at a bank that was a bus ride away. The QR code was a convenience. The convenience was the product of a series of decisions made, over fifteen years, by people the vendor had never met — the decision to build Aadhaar, the decision to establish the NPCI, the decision to build UPI, the decision to demonetize eighty-six percent of the country's currency overnight, the decision to make the rail free. The vendor did not know about these decisions. He knew about the tea.

The engineer, who worked for a technology company whose offices were visible from the vendor's cart, also did not know the institutional history. He knew that the payment had settled, that his bank account had been debited, and that the vendor's bank account had been credited. He knew that the transaction had been free, instant, and frictionless. He did not know that the system that processed his payment had not existed ten years earlier, or that it now processed more transactions than the two largest payments networks in the world combined. He knew about the tea. He drank it, returned to his office, and forgot about the transaction.

The transaction was unremarkable. The transaction was, from the perspective of both participants, ordinary. The ordinariness was the achievement. The system had become, for the vendor and the engineer, as routine as cash. That ordinariness had taken eight years to build.

* * *

In Singapore, a cryptocurrency trader settled a fifty-million-dollar position in Tether. He had purchased the Tethers on an exchange, paying with a wire transfer from his bank, and used them to settle a trade with a counterparty in Hong Kong. The transaction settled in minutes, on a blockchain that anyone could inspect but no single authority supervised. The Tethers were liabilities of Tether Limited, a company incorporated in the British Virgin Islands, whose reserves — substantially United States Treasury bills — made it one of the largest holders of American government debt in the world. The company had never published a full audit. It had been fined by the CFTC for misrepresenting its reserves. It continued to operate, to grow, and to issue dollar-denominated liabilities at a rate that no regulator had been able to constrain. The fifty million dollars the trader settled was, for Tether, a modest transaction. For the global dollar system, it was invisible.

The trader did not think about the company that issued the Tethers. He thought about the price, the spread, and the speed. The Tethers were a tool — a dollar substitute that moved at the speed of the blockchain, without the delays and the documentation of the banking system. The trader used Tethers because they were faster than dollars, cheaper than dollars, and easier to move than dollars. He did not use them because he believed in cryptocurrency, or because he distrusted the Federal Reserve, or because he had read the company's attestations. He used them because they worked.

The company that issued the Tethers employed roughly a hundred and fifty people. The Federal Reserve, which was responsible for the stability of the dollar system in which the company operated, employed roughly twenty thousand. The European Central Bank employed roughly three thousand five hundred. The company, with a hundred and fifty employees and a hundred and twenty billion dollars in dollar-denominated liabilities, had become one of the most important dollar institutions in the world. It had achieved this status without a banking license, without comprehensive supervision, and without, at any point in its history, publishing a full audit of its reserves. The trader did not know any of this. He knew about the price, the spread, and the speed.

* * *

In Guadalajara, a man sat in an office above a furniture store and matched five hundred thousand dollars, held by a courier in Los Angeles, with a Colombian importer's obligation to pay for electronics from a supplier in California. The man was a peso broker. His office contained a desk, a calculator, and a laptop running a spreadsheet that recorded the dollars available from drug sales in American cities and the pesos needed by importers in Latin America. He matched the two, instructed the courier in Los Angeles to deliver the cash to the supplier, and arranged for pesos to be delivered to the drug trafficking organization that had generated the dollars. The transaction settled in four hours. The fee was one to three percent. No bank was involved. No report was generated. The anti-money-laundering regime that the United States government maintains, at a cost of tens of billions of dollars a year, did not see it. It was one of thirty to forty billion dollars' worth that the Black Market Peso Exchange would process that

year.

The peso broker was forty-two years old. He had been in the business for fifteen years. He had seen colleagues arrested, networks dismantled, and routes disrupted. He had also seen the networks replaced, within weeks, by new networks that served the same clients, used the same routes, and operated on the same principles. He did not think of himself as a criminal. He thought of himself as a financier — a man who matched buyers with sellers, who cleared transactions, and who took a commission for the service. The fact that the buyers were drug trafficking organizations and the sellers were importers was, from his perspective, a detail of the clientele, not a detail of the business. The business was clearing. The clearing was efficient. The efficiency was the product of the absence of compliance — the absence of forms, of identity verification, of reporting, of the costs that the legitimate banking system bore and that the BMPE did not. The broker knew this. He did not dwell on it. He had a ledger to balance, and the courier in Los Angeles was waiting for instructions.

The courier in Los Angeles was twenty-six years old. He drove a rented sedan. He collected cash from small businesses — a car wash, a restaurant, a furniture store — that served as collection points for the proceeds of drug sales in the city. He did not know the peso broker in Guadalajara. He knew the representative who gave him his instructions, and the representative knew the broker. The chain was deliberately segmented — each participant knew only the person above and below them in the chain, so that the arrest of any single participant would not compromise the others. The courier collected the cash, delivered it to the designated recipient, and received his payment — a percentage of the amount collected, typically one or two percent. The work was steady. The work was illegal. The work was, from the courier's perspective, a job.

* * *

In Buenos Aires, a schoolteacher entered a clothing store on Avenida Santa Fe, asked for the manager, and was led through a door, down a hallway, into a back room where a man sat behind a desk with stacks of hundred-dollar bills. She earned four hundred thousand pesos a month — about four hundred dollars at the blue rate. She wanted to exchange fifty thousand pesos for dollars, to hold as savings, because she did not trust the peso and no one she knew trusted the peso. The man quoted the blue rate. She handed over the pesos. He counted out the dollars. The transaction took under five minutes. It was technically prohibited by the capital controls the central bank maintained. It was also universal — repeated by millions of Argentines, in thousands of back rooms, across a country that had learned to operate its monetary system without its central bank.

The schoolteacher had been doing this since 2002. She had been twenty-three when the corralito froze her bank account, and she had watched, in the devaluation that followed, as the savings she had accumulated — the equivalent of two thousand dollars, earned over three years of teaching — lost three-quarters of their value. She had learned the lesson that every Argentine of her generation learned:

do not hold pesos. Hold dollars. Hold dollars in cash, under the mattress, in a safe deposit box, in the back room of a clothing store. The lesson had governed her financial behavior for twenty-three years. She had never, in those twenty-three years, held a significant sum in pesos for longer than the time it took to convert them to dollars. She had taught her daughter the same lesson. Her daughter, who was sixteen, had never seen her mother deposit money in a bank. She had seen her mother go to the clothing store.

The man behind the desk in the back room was fifty years old. He had been operating the cueva for twelve years. He published, each morning, a rate that reflected the actual supply and demand for dollars in the Buenos Aires market. The rate was acknowledged by every participant in the economy — by the households that checked it on their phones, by the businesses that used it to price their contracts, by the economists who cited it in their analyses, and by the central bank itself, whose officials consulted the blue rate, in private, to assess the gap between their published fiction and the market's reality. The man did not set the rate. The market set the rate. He was a price-taker, not a price-maker. The central bank, which published a different price, was a price-pretender. The man knew this. The schoolteacher knew this. The central bank knew this. Everyone in Argentina knew this.

* * *

In London, a trader at a hedge fund monitored his nickel positions on the London Metal Exchange. The Ring — the circle of red leather benches where traders have shouted bids and offers at each other since 1877 — was conducting its twice-daily session. The electronic system, which handles the vast majority of the exchange's volume, was executing orders at microsecond speed. The trader's risk models had been updated since March 2022, when the exchange cancelled four billion dollars of trades to prevent its clearinghouse from failing. The updates were adequate. They were also, like all updates, a response to the last crisis.

The trader had been in the market on March 8, 2022. He had watched the price of nickel go from forty-eight thousand dollars a tonne to over a hundred thousand in a matter of hours. He had watched the exchange suspend trading. He had watched the exchange cancel the trades — retroactively voiding four billion dollars of transactions, declaring that its own price-discovery mechanism had produced prices that were invalid. He had not been one of the traders whose trades were cancelled. He had been on the right side of the move. But he had watched colleagues — traders at Jane Street, at Elliott Management — lose hundreds of millions of dollars in profits that the exchange had erased with a press release. He had also watched the exchange survive, which was the point of the cancellation. If the trades had stood, the clearinghouse would have failed. If the clearinghouse had failed, every participant in the market would have taken losses. The exchange had chosen its own survival over the integrity of its prices. The trader understood the choice. He did not have to like it.

He monitored his positions on the screen. The risk models were updated. The procedures were current. The procedures were also, by the nature of the system they governed, one step behind the trades they were designed to monitor.

* * *

In La Hulpe, a suburb of Brussels, a SWIFT message traveled from a bank in Frankfurt to a bank in Mumbai. It was one of roughly forty-five million the network would process that day. The message was a formatted text — a structured set of fields specifying the sender, the receiver, the amount, the currency, and the purpose of the payment. The message did not carry the money. It carried instructions for the money. The money itself moved through correspondent banking accounts — separate arrangements between the banks. SWIFT was the messaging layer. The settlement was separate.

The building in La Hulpe — a low-rise office complex surrounded by parking lots and landscaped berms, marked by a modest sign bearing the SWIFT logo — looked exactly as it had looked before February 2022. It was no longer what it had been. It had been used, once, as a weapon. Seven Russian banks had been cut off from the network, at the direction of the Western governments, in response to the invasion of Ukraine. The weapon had been deployed in a weekend. The cooperative that maintained the network had complied, because it was subject to EU law and had no choice. The neutrality that the cooperative had maintained for forty-nine years — the fiction that it was a utility, like the postal service, and should not be used as an instrument of foreign policy — had been retired in forty-eight hours. The fiction had not returned. The building looked the same. The function had changed.

The message from Frankfurt to Mumbai passed through the switching infrastructure in La Hulpe without incident. It was processed, routed, and delivered, in the same manner as the other forty-five million messages that would be processed that day. The system worked. The system had always worked. The system's problem was not that it did not work. The system's problem was that it had been revealed to be something other than what it claimed to be.

* * *

In Jersey City, a market-making firm's systems executed thousands of trades per second on the U.S. equity exchanges. The risk management team monitored the positions on screens, alert to the possibility of a malfunction — a bug, a misconfiguration, a piece of retired code reactivated by a deployment error — that could, in minutes, produce losses the firm's capital could not absorb. The firm's procedures had been updated since August 2012, when Knight Capital lost four hundred and forty million dollars in

forty-five minutes because a flag in a directory reactivated a function that had been retired in 2003. The procedures were current. They were also, by the nature of the system they governed, one step behind the trades they were designed to monitor.

The firm's chief technology officer had, earlier that morning, signed off on a deployment — a routine update to the firm's order-routing software. The deployment had been tested. The tests had passed. The deployment had been reviewed by a team of engineers, by a risk officer, and by the CTO himself. The deployment was, by every available measure, correct. The CTO knew, however, that the deployment was correct only in the sense that the tests had not detected an error. The tests could not test for errors they were not designed to detect. The Power Peg code that had killed Knight Capital had not been detected by Knight's tests, because Knight's tests had not been designed to detect the reactivation of retired code. The CTO knew this. He signed off on the deployment anyway, because signing off on deployments was his job, and because the alternative — not deploying, ever, for fear of the undetectable — was not an alternative that a market-making firm could afford. The deployment went live. The systems executed. The trades flowed. Nothing went wrong.

* * *

In Bonn and Frankfurt, BaFin — the German federal financial supervisor — carried the memory of Wirecard. The agency's president had resigned. Its procedures had been updated. Its officials reviewed market data, examined filings, and conducted the routine supervision their mandate required. The routine was competent. It was also incomplete, because the agency's supervision extends to the entities it is authorized to supervise — the banks, the financial services institutions, the insurance companies — and not to the parallel rails that operate outside its jurisdiction. The officials are aware of the parallel rails. They cannot, under the legal framework that authorizes their supervision, supervise them.

In Munich, Markus Braun sat in a jail cell. He had been there since 2020. His trial had been delayed repeatedly by the complexity of the case and by the legal maneuvers of his defense team. His former chief operating officer, Jan Marsalek, remained at large — reportedly in Moscow, reportedly under Russian state protection, reportedly alive. The reports could not be independently confirmed. The €1.9 billion that Wirecard had admitted, in June 2020, probably did not exist had never been found. It had never existed. The fiction had been maintained, through years of audit opinions and regulatory filings, until the fiction could no longer be maintained. The fiction was now, for Braun, a criminal case. For Marsalek, a life in exile. For the investors who had lost their money, a loss. For BaFin, a memory that the agency's officials carried, every morning, into their offices.

* * *

In Washington, D.C., a consumer bought a four-dollar-and-fifty-cent coffee with a Chase Sapphire Reserve card. The merchant paid roughly thirteen cents in interchange to the issuing bank. The consumer received roughly thirteen and a half cents in rewards points. The net transfer — half a cent, from the merchant to the consumer — was invisible, as it has been for fifty years, embedded in the price of the coffee, unitemized on the receipt, and maintained by a duopoly that has defended its two percent through five decades of attempted reform.

The consumer did not know about the interchange fee. She knew about the rewards. The rewards were the reason she carried the card. The card offered three points per dollar on travel and dining, a sign-up bonus worth roughly fifteen hundred dollars, and a set of perks — lounge access, travel credits, insurance — that she valued at more than the annual fee. The rewards were funded by the interchange. The interchange was paid by the merchant. The merchant passed the cost to all consumers, in the form of higher prices. The consumers who paid with cash — who tended to be lower-income — paid the same higher prices, but received no rewards. The transfer, from cash payers to rewards card holders, was invisible. It had been invisible for fifty years. The consumer did not know about the transfer. She knew about the coffee and the points.

The merchant knew about the interchange. He paid it, every month, in the fees that his payment processor charged him. He could not refuse the card. The network rules required him to accept all Visa cards, including the premium cards with the higher interchange. His only option, if he wished to avoid the interchange, was to stop accepting Visa entirely, which would cost him customers. He accepted the card. He paid the fee. He raised the price of the coffee, slightly, to cover the cost. The consumer paid the higher price. The consumer received the rewards. The system continued.

* * *

In the village near Lake Victoria, the two thousand shillings arrived. The mother, notified by a text message on her Nokia, walked to the M-Pesa agent — a shop that sold airtime, sugar, and cooking fat — and presented her phone number. The agent, a woman roughly the mother's age, confirmed the transaction and counted out two thousand shillings in cash. The mother folded the notes, placed them in the pocket of her dress, and walked home.

She bought flour from a neighbor who sold it from a sack in her kitchen. She bought cooking oil from another shop. She bought a prepaid electricity token from the agent, entered it into her meter, and the lights came on. The two thousand shillings that had been in Nairobi that morning — ten seconds, three hundred kilometers, thirty shillings in fees — had been absorbed into the village economy by afternoon. The flour was in the kitchen. The oil was on the shelf. The electricity was on. The money had done what money does.

* * *

None of the transactions described in these pages passed through the system that appears in the textbooks. The textbooks describe a system of banks, regulated by central banks, clearing payments through supervised infrastructure. That system exists. It handles the payrolls, the corporate wire transfers, the bond coupons, and the dividend payments of the formal economy. It is the system that the regulators regulate, that the auditors audit, and that the economists model.

The transactions in this chapter passed through a different system, or rather through a collection of different systems, none of which was designed as a whole and none of which is supervised as a whole. A telecom in Nairobi. A non-profit in Mumbai. A company in the British Virgin Islands. A broker in Guadalajara. A back room in Buenos Aires. A cooperative in Belgium. An exchange in London. A server in New Jersey. Each was built to solve a specific problem — to move a specific kind of money, for a specific kind of user, at a speed and a cost that the formal system could not match. Each was built by people who were not coordinating with each other. Each grew, through use, into a piece of infrastructure that its builders did not entirely anticipate.

The pieces are not connected to each other, except by the fact that they all move money. They do not share a common architecture, a common regulatory framework, or a common set of participants. They include the legal and the illegal, the public and the private, the audited and the unaudited, the beneficial and the destructive. They include systems built by central banks and systems built by criminal organizations. They include systems that charge nothing and systems that extract two percent. They include systems that settle in seconds and systems that settle in days. The only thing they share is that they all work — the money arrives — and that none of them was built by the institutions that were, formally, responsible for building the things that move money.

The institutions that were responsible — the banks, the central banks, the regulators, the clearinghouses — were designed for a world in which money moved at a speed they could match. Money no longer moves at that speed. The gap between the speed of the money and the speed of the institutions is the gap in which the collection of systems described in this book has grown. The gap is not closing.

* * *

The mother, who is sixty-three, sits on a chair outside her house in the late afternoon sun and drinks a cup of tea. The tea was brewed with the flour and the cooking oil and the electricity that the two thousand shillings purchased. The money has been absorbed into the village economy. The flour was

bought from a neighbor. The oil was bought from a shop. The electricity token was bought from the agent. The money has done what money does.

Her phone, propped against the tea pot, chimes. Five hundred shillings, from a son in Mombasa, sent three minutes ago. The mother looks at the screen, notes the amount, and finishes her tea. She will collect it tomorrow, or the next day. The money will wait.

The neighbor calls out from across the path. The mother calls back. They talk about the price of flour, the health of a child, the prospect of rain. The phone sits against the tea pot, the screen dark, the five hundred shillings held in it — waiting, as the money in these systems always waits, for someone to come and collect it.

The money is in no bank. The money is in no clearinghouse. The money is in no regulated institution. The money is in a phone, propped against a tea pot, in a village three hundred kilometers from the nearest central bank, on a system built by a telecom, regulated under a category that did not exist when the system was built, serving a woman who has never had a bank account and who does not need one. The money is there because the system works. The system works because the institutions that were supposed to build it did not. The institutions did not build it because they could not match the speed at which the money needed to move.

The mother finishes her tea. The sun moves across the yard. The phone waits.