WHEN THE ALGORITHM GETS IT WRONG

When the Algorithm Gets It Wrong

AI harms and ethical questions from the AIAAIC record

What happened in hiring and welfare, courts and roads, classrooms and workplaces

Kim Kyung-jin, Attorney at Law

Table of Contents

CHAPTER 1

Bias and Discrimination: Human Prejudice Written into Code

1.1

Algorithmic Bias in Hiring, Finance, and Services

The rejection emails always arrived at dawn.

Derek Mobley was a Black man over forty, and suffered from an anxiety disorder and depression. He worked in the information technology field, and had certifications and experience. He submitted applications to about a hundred companies that used a hiring system from a company called Workday.

He was rejected about a hundred times.

Looking at the times the rejection emails arrived, one could tell that they were not sent by a person. One in the morning, three in the morning. Human hiring managers sleep at those hours. Machines do not sleep. They do not drink coffee, and they do not feel sorry. They open a document and close it in three seconds.

Mobley filed a lawsuit. His claim was that Workday's artificial intelligence acted like a sieve that filtered out age, race, and disability. The court viewed that Workday trained its artificial intelligence with existing employee data and used protected characteristics as proxy indicators, and approved a nationwide class action lawsuit.

In March 2026, the court rejected Workday's argument that the Age Discrimination in Employment Act did not apply to job applicants. On June 16, a ruling was issued that Workday could be held liable even when used by employers outside California. The number of rejections covered by this lawsuit is 1.1 billion.

That is 1.1 billion dawn emails.

The reason this story is scary is that there are no villains. Nobody gathered in a conference room and resolved to reject people over forty. They simply fed the data of people hired over the past ten years into a machine. The machine diligently learned that pattern. And it diligently repeated it.

The term machine learning sounds difficult, but its principle is similar to a primary school math class. It involves showing a ton of answers without providing the rules, and telling it to find the rules from them. If the past answers are biased, the machine learns the bias as a rule. From the machine's perspective, it did a good job. It did as it was told.

Ten years ago, Amazon learned that fact first.

In 2014, a secret project began at the Amazon development center in Edinburgh, Scotland. It was a system that gave star ratings when resumes were submitted. From one star to five. It essentially gave people the star ratings that are given to products.

The developers inputted ten years' worth of technical job applications as training data. The technical staff Amazon hired during those ten years were mostly men. The machine made a conclusion. Men are better applicants.

When the machine found the word women in a resume, it deducted points. A resume that said president of the women's chess club was penalized. The scores of applicants who graduated from women's colleges were also lowered entirely. Conversely, if verbs frequently used by male engineers, such as executed or captured, were included, the score went up.

Amazon tried to fix this bias. It failed. The project was scrapped in 2017. Fortunately, Amazon stated that the system was not used in actual hiring.

The problem is that other companies kept making similar things.

Kyle Behm was a student at Vanderbilt University. He took a temporary leave from school to treat his bipolar disorder, and applied to local stores to earn living expenses. Kroger Supermarket, Home Depot, Walgreens, PetSmart. They were part-time cashier jobs.

He was rejected from all of them. He could not even get to an interview.

This was because an online personality test called Unicru, made by a company named Kronos, turned on a red light for Kyle. The test detected Kyle's mental health history, and determined that such a person is highly likely to ignore customers when they get angry.

Kyle's father, Roland Behm, was a lawyer. In 2014, he filed a complaint with the Equal Employment Opportunity Commission, and filed a lawsuit against the companies for discriminating against people with mental disabilities.

Kyle ended his own life in 2019. It was before the lawsuit was finished.

The fact that there was a young man who could not stand at a checkout counter because he failed a personality test. The fact that what made that judgment was not a person but a webpage. And the fact that the webpage had no channel for complaints. Reading these three sentences together sends a chill down your spine.

iTutorGroup's system was even more explicit. This online education company's automated hiring program automatically eliminated women aged fifty-five or older and men aged sixty or older. Over two hundred teachers were cut because of their age despite being qualified.

The way it was exposed is comical. An applicant applied with her real birthdate and was rejected. However, when she applied again with only her age lowered, she was immediately contacted for an interview. A person would have blushed, but the system was unfazed. The Equal Employment Opportunity Commission imposed a settlement of 365,000 dollars.

People thought it would get better when generative artificial intelligence came out.

In 2024, Bloomberg researchers gave GPT-3.5 a ton of fake resumes with the exact same qualifications, only changing the names. They were names from which race and gender could be guessed. The results were as follows. In technical job screenings, names guessed to be Black women were selected 36 percent less. Female names were pushed to human resources management roles.

In 2021, the German broadcasting station Bayerischer Rundfunk tested an artificial intelligence competency assessment from a startup called Retorio. When the same person said the same things, the score changed if they wore glasses. It changed again if they wore a hijab. A judgment came out that they looked diligent if a bookshelf was visible behind them.

It is funny that setting up a bookshelf improves one's personality. The laughter stops at the point that someone got a job and someone else did not because of that judgment.

The same thing happened in the teaching profession. In 2012, Sarah Wysocki, a fifth-grade teacher at a Washington, D.C. public school, was fired by a teacher evaluation system called IMPACT. The evaluations from her fellow teachers and parents were excellent. However, there was test cheating at a school the children previously attended, and the inflated scores from that dragged down Wysocki's value-added score.

Two hundred and five teachers lost their jobs due to a single unverified formula. What that formula was has not been disclosed.

Algorithms that handle money have worked longer and more quietly.

A 2018 report from researchers at UC Berkeley examined thirty-year fixed-rate mortgages guaranteed by Fannie Mae and Freddie Mac from 2008 to 2015. At

that time, there was talk that discrimination would disappear if lending decisions were handed over from people to algorithms.

The numbers told a different story. Black and Latino borrowers paid 5.6 to 8.6 basis points more on purchase loans and 3 basis points more on refinances. A basis point is a financial unit equal to 0.01 percent. It looks like a small number.

When this small number was added up across the entire United States, it totaled 250 million to 500 million dollars annually.

The reason is even more bitter. Lenders used artificial intelligence to discover that minority borrowers lacked the time and information to compare multiple products. They then charged those people higher prices. This is called strategic pricing. The additional margin they captured this way was 11 to 17 percent.

The investigative news outlet The Markup and Yonhap News analyzed federal housing loan data together. A secret approval algorithm based on classic FICO scores rejected applicants of color 40 to 80 percent more often than white applicants. In some major cities, it exceeded 250 percent.

Laura Blattner and Scott Nelson of Stanford Business School went even deeper. They showed that problems occur even without bad intentions in the algorithm. Low-income people and minorities have thin credit histories. When the history is thin, predictive accuracy drops 5 to 10 percent. One or two instances of delinquency from long ago can destroy an entire score.

If you are poor, there is less data; if there is less data, the machine makes mistakes; if it makes mistakes, you are classified as a risky person. And that classification is labeled as objective credit risk.

At Wells Fargo, a single calculation error drove people onto the streets. From 2010 to 2015, this bank's automatic mortgage modification review software contained an error that miscalculated attorney fees. For over five years, no one

noticed.

Eight hundred seventy people who were eligible to receive modified loan terms received rejection notices. At least five hundred forty-five of them had their homes foreclosed. The bank fixed the error at the end of 2015 and did not disclose this fact for several years. The compensation later determined was eight million dollars.

When you consider the value of one house, the amount does not add up.

Insurance was the same. Risk assessment algorithms used by GEICO, Safeco, and Nationwide charged residents of neighborhoods with large minority populations up to 30 percent more in premiums, even when the accident risk was the same. Allstate went even further. When it seemed a customer would not leave, it raised the price. Loyal, long-time customers paid more. This was detected by regulators.

Even after entering the workplace, algorithms follow you.

In 2021, the Bologna Labor Court in Italy ruled that the Frank algorithm used by delivery platform Deliveroo was discriminatory. Frank scored delivery workers' reliability and engagement. If workers did not show up for scheduled shifts, their score was deducted. It did not ask why.

The score was deducted even if workers were sick and could not come in. It was deducted even if they participated in strikes protected by law. If the score was deducted, they could not get good delivery assignments next time. The right to strike remained in the law, but the code contained a structure where striking meant going hungry.

An automated background check system operated by Checkr screened out Uber, Lyft, and Postmates drivers in large numbers starting in 2015. Criminal records of other people with the same name, records already past the statute of limitations, remained attached. Drivers whose accounts were suddenly

blocked did not know where to file a complaint.

Professor Veena Dubal of UC Hastings Law School gave this a name. Algorithmic wage discrimination. Uber, Lyft, and Amazon collect worker data and pay different amounts for the same work. Because workers do not know what the person next to them earns, they cannot compare.

Passengers did not have it easy either. According to research, the pricing and dispatch algorithms of Uber and Lyft imposed higher fares and longer wait times on areas with large minority populations.

Facebook's advertising system filtered job postings by age. Verizon, Amazon, and Goldman Sachs used age-targeting features to make postings invisible to older job seekers. A 2021 Northeastern University study and a 2025 French Human Rights Ombudsman ruling went a step further. Even when advertisers did not specify gender, Facebook's internal delivery algorithm automatically showed different job postings to men and women.

The price of things was also determined by looking at the person.

Airbnb's Smart Pricing is a feature that automatically adjusts accommodation rates. When researchers led by Professor Param Vir Singh of Carnegie Mellon University examined it, they found that white hosts used this feature more often, and as a result, the income gap between white and Black hosts actually widened.

Princeton Review's online tutoring fees varied by zip code. Higher prices were charged in areas with large Asian American populations. The media called this the Tiger Mom tax.

The travel site Orbitz displayed more expensive hotels at the top for people using Mac computers. This was because statistics showed that Mac users had higher average incomes. It assumed that if a laptop had an apple on the lid, the wallet was thicker.

Amazon had a secret algorithm called Project Nessie. It was an experiment where the company slightly raised prices for its own products and watched to see if competitors would follow. The Buy Box algorithm showed consumers products that paid Amazon higher fees instead of the cheapest product.

All these cases have one thing in common. When the problem surfaced, what the companies said was almost identical.

It is the result of the algorithm learning on its own. It is a technical error. The internal structure is a trade secret and cannot be disclosed.

Three sentences were enough. The people who selected the data, who set the objective function, and who benefited from the results were all in the conference room, yet the responsibility went to the code. Code does not make excuses. It does not apologize. It does not appear in court.

There is no more convenient employee in the world.

1.2

Algorithmic Malfunctions in Welfare Benefits and the Social Safety Net

The letter looked ordinary. It was in an Australian government envelope, with a government font, and a government seal was stamped on it. Inside, it read like this. You have been overpaid in welfare benefits. Please repay the amount below.

The recipients were surprised twice when they saw the amount. Some letters had amounts exceeding ten million won. And they were surprised a third time. There was no explanation as to why they had to repay it.

More than four hundred thousand people received this letter.

This was the work of the Online Compliance Intervention system, a program run by the Australian Department of Social Services and Centrelink from 2016 to 2019, which people called Robodebt. Robo means robot, and debt means debt. It means a debt created by a robot.

The mistake committed by this system was at the level of elementary school arithmetic.

The government matched the annual income data held by the National Tax Service with welfare benefit records. The problem was that the time units of the two data sets were different. The National Tax Service had a year's worth of data as a single lump sum, and the welfare benefits went out once every two weeks.

The system chose the easy way. It divided the annual income by fifty-two and assumed that the person earned the same amount of money every week.

This would be a correct calculation for a full-time office worker. However, people receiving welfare benefits are generally not full-time workers. They are people who work for three months during the strawberry picking season and rest the rest of the time, people who work two jobs this month and cannot work a single job next month, and people who reapply for benefits every time their contract ends.

What happens if you divide the annual income of such a person by fifty-two? It is calculated that they earned money even in the weeks they did not work. Therefore, the conclusion is drawn that they received benefits unfairly.

Debt collection notices went to people who had no debt.

To raise an objection, they had to bring pay slips from several years ago. If they did not have them, it became a debt they owed. It was a structure where citizens had to prove their innocence, rather than the government proving its calculation. Debt collection agencies got involved. Tax refunds were seized.

Some people could not endure it. There were people who took their own lives, unable to find a place to resolve their unfairness.

At the Royal Commission hearings in 2023, a truly terrifying fact came out. There had been multiple warnings that this system was illegal and that its calculation method was flawed. Legal experts had warned about it, and internal employees had warned about it.

High-ranking public officials and ministers knew. And they kept running it.

This was because they needed the achievement of having saved the budget. If a person checks each case one by one, it takes time and costs money. Eliminating that verification process was the core function of this system. The conclusion was a settlement of 1.8 billion Australian dollars and an official apology from the government.

The apology did not reach the deceased.

In India, fingers were the problem.

The Indian government manages welfare with a biometric authentication system called Aadhaar. You must scan your fingerprint to receive food rations. In a remote village in the state of Jharkhand, this method killed people.

The fingerprint reader kept failing. The fingertips of a person who has done farm work and manual labor all their life are worn down. Their fingerprints become faded. Moreover, the internet does not reach mountain villages well. A whole day passes while waiting for a signal.

Marginalized residents, including the Parihaiya, were denied their legally guaranteed food rations. People who starved to death or died of malnutrition emerged in several villages.

It was not that the machine failed to recognize the person. The rule that dictated not giving food if the machine could not recognize them killed people.

In the state of Haryana, a system called Parivar Pehchan Patra processed thousands of living people as deceased. If you are recorded as dead, your pension is cut off and your food rations are cut off. When investigated later, 70 percent of the suspended pensions were due to the misjudgment of the algorithm.

There were elderly people who had to go to government offices to prove they were alive.

Samagra Vedika in the state of Telangana incorrectly predicted income and employment status, quietly eliminating the food cards of fifteen thousand households. They were restored only after people rose up.

In Jordan, a poverty measurement algorithm called Takaful, supported by the World Bank, was run. It scored several characteristics of households to determine the targets for support. However, truly starving households were filtered out for falling short of the score. Those who tried to measure poverty

with numbers failed to recognize poverty.

Brazil's Mep INSS app was created to reduce the waiting lines for pension applications. As a result, the lines did shrink. This was because applications were automatically rejected. If there was a typo on the form or a question was misunderstood, they were eliminated just like that. Elderly people in rural areas who were not familiar with smartphones were caught. To apply again, they had to hire a lawyer.

It was no different for wealthy countries.

The United Kingdom's Universal Credit is a system that combined several welfare benefits into one. The calculation algorithm that went into this determined the benefit based on the money that came in during a single calendar month. Human Rights Watch uncovered this structure in 2020.

Some workers receive their wages once every four weeks. Then, in certain months, the wages come in twice. The system judged that this person had suddenly become rich. Their benefits for that month were drastically cut or did not come out at all.

Money for food disappeared depending on what day of the week the calendar started on.

The fraud detection algorithm of the UK's Department for Work and Pensions operated more quietly. A model called Advances and a housing benefit surveillance algorithm looked at the recipient's age, nationality, disability status, and marital status to assign a risk score.

A list of people who received high scores came out. They were low-wage, part-time female workers of Bulgarian and Polish nationality, and people with disabilities.

Two hundred thousand people were classified as a high-risk group and had their support stopped or were investigated over three years. More than

two-thirds of them were legitimate recipients who had done nothing wrong.

Denmark's welfare agency UDK ran more than sixty fraud detection algorithms. It gathered residential status, nationality, family relationships, medical records, and even departure histories. The state gathered where they had been, who they lived with, and what illnesses they had all in one place and scored them.

International human rights organizations saw this as a social credit system banned by the European Union's Artificial Intelligence Act. Amnesty International released a report titled "Code as Injustice." It documented how people with disabilities, low-income populations, immigrants, refugees, and racial minorities faced high risk of being marked for investigation.

When Amnesty asked Danish authorities to show the code and data, they refused.

The Swedish Social Insurance Agency ran a fraud prediction model starting in 2013. When the Lighthouse Report examined it, women, immigrants, people with foreign backgrounds, low-income earners, and people without a university degree were flagged more often.

The French social security organization CNAF's risk assessment algorithm scored almost half the population. Single-parent households, low-income populations, and people with disabilities were automatically marked as suspects. It did not explain why the scores came out that way. Instead, investigators came to search their homes.

In the Netherlands, the tax authority launched a machine learning algorithm from 2013 to 2019 to catch childcare subsidy fraud. Thousands of parents were labeled fraudsters. Dual citizenship was used as a risk signal. Rotterdam's system categorized people by ethnicity, age, sex, and whether they had children.

There was also an algorithm made to protect children.

Allegheny County in Pennsylvania used a family screening tool to predict child abuse risk. When a report came in, the tool would score it. If the score was high, an investigator went to the home.

Researchers from Carnegie Mellon University analyzed the results. Two-thirds of Black children reported were recommended for investigation. Half of White children were.

The reason lay in what the tool measured. The formula included a history of receiving government assistance. It also included mental health diagnosis records.

Receive support because you are poor, your risk score rises. Receive counseling because you are sick, your risk score rises. Seeking help becomes grounds for suspicion.

A scale where more help received means higher chance of losing your child. You cannot create such a scale and call it fair.

An Oregon system modeled after this tool was shut down in 2022 after racial bias was discovered. The Eckerd Rapid Safety Feedback used in Illinois was also discontinued due to data errors and inflated performance claims.

Denmark's child protection agency's decision support system showed strange patterns when tested by Copenhagen IT University. With identical situations, different ages produced different risk scores.

Saying it would save 700 million Australian dollars in disability support funding, the Australian government tried to introduce Robodebt, an automated assessment tool. It compressed disabled people's lives into a few standard categories, then calculated needed support. Blind organizations and disability groups strongly opposed it, and the plan was abandoned.

In hospitals, algorithms determined discharge dates.

United Healthcare Group and Humana, America's two largest health insurers, used nH Predict to forecast how many days of rehabilitation therapy elderly and disabled patients needed. Even if a doctor said more was necessary, when the system's scheduled discharge date arrived, insurance coverage stopped.

In a lawsuit, this system's claim denial error rate was revealed. Ninety percent.

Deny ten times, and nine times you are wrong. Yet it kept being used. The reason was simple. Most patients who were denied did not appeal. For elderly and sick people, fighting an insurance company is difficult.

In November 2021, eighty-six-year-old Joan Barros fell and broke her leg. Medical staff said physical therapy was absolutely necessary. When the algorithm's scheduled discharge date came, coverage for treatment stopped.

UnitedHealth's alert system illegally restricted mental health treatment approvals and was banned by several state governments. Cigna's PxDx was a system that allowed doctors to mass deny claims with a single click without opening a patient's chart. Evernorth's DR algorithm increased the pre-authorization denial rate to as high as 20 percent.

Deloitte-created TIERS in Texas and TEDS in Tennessee caused hundreds of thousands of low-income people and people with disabilities to lose Medicaid eligibility due to program errors and misdirected addresses. Notices went to wrong addresses. With no reply, eligibility was denied.

Counting the countries mentioned: Australia, India, Jordan, Brazil, United Kingdom, Denmark, Sweden, France, Netherlands, United States. Different systems. Different languages.

But the people caught are remarkably similar. Poor, old, sick, immigrants, people with nowhere to speak.

Not coincidence. These systems were introduced to save money. To save money, you must cut payments. To cut payments, you must cut people. The safest person to cut is someone who cannot object.

Algorithms excel at that calculation.

1.3

Facial Recognition Technology's Misidentification of People of Color and Wrongful Arrest

Two- and five-year-old daughters were watching from the yard.

On January 9, 2020, in the front yard of a house in Farmington Hills, Michigan, Robert Williams had just returned home from work. Detroit police burst in and handcuffed his wrists in front of his wife and children.

The charge was watch theft. It was the case of five watches worth $1,800 disappearing from a Shinola store in Detroit in October 2018.

The evidence was this. There was blurry footage captured by the store's surveillance camera. A Michigan State Police examiner captured a frame from that footage and entered it into a facial recognition database. The system produced Williams's old driver's license photo as a candidate.

What comes next is important. A store security officer who had not even been at the crime scene received several photographs of Black men and selected Williams.

Williams spent thirty hours in custody.

In the interrogation room, a detective placed a photograph from the surveillance camera on the desk and asked, "Is this you?" Williams held the photo next to his face and said, "This person is not me. Do you think all Black people look the same?"

There is a backstory to this case. At the time, Detroit Police Chief James Craig already knew that the technology had a 96 percent error rate in identifying suspects.

They went to someone's front yard carrying a tool that was wrong ninety-six times out of a hundred.

Robert Williams became the first official record of wrongful arrest due to facial recognition error in the United States. In June 2024, Detroit City settled with the American Civil Liberties Union for $300,000 and changed regulations so that arrest warrants could not be obtained based solely on facial recognition leads and photo comparison.

As of 2026, more than twelve people are known to have been arrested this way. That is only the known number.

A new lawsuit was also filed on June 10, 2026. Robert Dylan of Florida was arrested in 2024. The basis was blurry surveillance footage and a 93 percent match result. The warrant application was missing evidence that he could not have been the perpetrator.

The number 93 percent captivates people. It sounds like a test score. But this number does not mean there is a 93 percent probability that this person is the perpetrator. It means this photo looks most similar among the millions of images in the database.

The most similar-looking person always exists. Even if the perpetrator is not among them.

Angela Lipps's story shows what happens when you don't know this difference.

July 14, 2025, Tennessee. Lipps, mother of five children and grandmother of five grandchildren, was watching her grandchildren at home. Federal marshals came in with guns drawn. They said she was a fugitive in a bank fraud case that occurred in Fargo, North Dakota.

Lipps had never been to North Dakota. She had never even been on an airplane. She had barely ever left more than 160 kilometers from her home.

That was the extent of the investigation conducted by detectives at the Fargo Police Department. They received facial recognition results, found photos on social media, and compared them by eye. They did not make a single phone call.

Lipps was transferred to North Dakota and was detained for 108 days. She was only released after her lawyer submitted bank transaction records proving she was in Tennessee at that time.

In the meantime, she lost her home. She lost her car. She even lost her pet dog.

The same system used by Detroit police continued to bite people.

In July 2019, Michael Oliver was arrested on charges of stealing a teacher's cell phone. A program from DataWorks Plus connected his photo in the police database with the perpetrator.

In February 2023, Porsha Woodruff was arrested on charges of robbery and carjacking. She was eight months pregnant. She experienced labor pains on the floor of the holding cell.

The investigator ran the surveillance footage through an algorithm and obtained Woodruff's driver's license photo from eight years earlier. He did not verify the physical characteristics of the wanted subject. The person captured on the surveillance camera was not pregnant.

You do not need artificial intelligence to distinguish between a person with a belly like Namsan Mountain and a person who does not have one. You need eyes. But the procedure of looking with eyes was missing.

In April 2025, Travis Williams was arrested on suspicion of indecent exposure in New York and was detained for more than two days. At that time, he was in Brooklyn, nineteen kilometers away.

Comparing physical characteristics makes you laugh. The actual perpetrator was 167 centimeters tall and weighed 72 kilograms. Williams is 188 centimeters tall and weighs 104 kilograms.

The police refused to verify location data or check with his employer.

In November 2022, Georgia resident Randal Quran Reid was accused of being the perpetrator who stole a luxury handbag in Louisiana and was detained for six days. A warrant was issued based on a single match result from Clearview AI. Reid said he was in Georgia at that time, but no one verified it. Later, the case was quietly dismissed.

Why do only these faces keep getting mistaken? The answer was already out there in 2018.

Joy Buolamwini and Timnit Gebru from MIT and Stanford released a report called Gender Shades. It was the result of testing facial analysis programs from IBM, Face++, and Microsoft.

For lighter-skinned men, the error rate in determining gender was less than 0.8 percent. For darker-skinned women, the error rate reached 34.7 percent.

That is approximately one in three.

The cause lies in the training data. Famous datasets like ImageNet, LFW, and BDD100K overwhelmingly contained photographs of white men. Machines distinguish faces they have seen frequently well. Faces they have seen less often they lump together.

The same pattern appeared when Georgia Institute of Technology analyzed the autonomous vehicle dataset BDD100K. The accuracy in recognizing dark-skinned pedestrians was on average 4.8 percent lower, and up to 12 percent lower at maximum.

This is saying that cars cannot recognize people. This passage will come up again later.

The warning was sufficient. The sales continued.

Clearview AI scraped three billion facial photos of people from Facebook, Instagram, Twitter, and LinkedIn without consent to create a search database. Platform companies demanded a halt and an Illinois Biometric Information Privacy Act violation lawsuit was filed, but the company sold its search services to investigative agencies worldwide.

We do not know which police station computer holds the graduation photo we uploaded ten years ago.

Amazon also tried to sell a product called Rekognition to the police and Immigration and Customs Enforcement. When the American Civil Liberties Union tested it in July 2018, it found that the faces of twenty-eight federal representatives and senators matched criminal photos. The proportion of people of color among the misidentified lawmakers was exceptionally high.

In October 2019, twenty-seven out of one hundred eighty-eight professional sports players in the New England region appeared as criminals.

Amazon said it guided users to set the confidence threshold at 95 percent or higher. That threshold was not followed in the field. People who read manuals are always a minority.

There was also exaggerated advertising. The Federal Trade Commission took issue with a company named Intellivision advertising that its facial recognition software had no gender or racial bias and had the world's highest accuracy. It said it trained with millions of photos, but in reality, it was a dataset of one hundred thousand modified photos.

Retailer Rite Aid intensively installed facial recognition cameras from 2012 to 2020 in stores in neighborhoods where many low-income earners and people

of color live. Thousands of customers were falsely accused as thieves and were kicked out of stores or searched. The Federal Trade Commission banned them from using this technology for five years.

Up to this point, these are at least stories of people who came back alive.

In a January 2022 Sunglass Hut robbery in Houston, Texas, Macy's and EssilorLuxottica's facial recognition system identified Harvey Murphy Jr. as the culprit. He was in Sacramento, California at that time.

While he was arrested in October 2023 and held in a detention center for two weeks, Murphy was gang-assaulted and sexually assaulted by three inmates. He was left with a permanent physical disability.

In Ponddale, Missouri, there was a train platform assault incident in 2021. The suspect had their face covered with a mask and clothes. The detective ran that photo through an algorithm.

Even if you input a covered face, the system presents candidates. This is because there is no function to answer that it does not know. Among the candidates produced that way was a twenty-nine-year-old father of four with no criminal record, Christopher Gatlin.

Police internal guidelines also stated that they must not arrest based solely on facial recognition results. The detective created a photo lineup, and Gatlin was in prison for over two years.

Kimberly Williams of Oklahoma was locked up for six months starting in June 2021. Detectives copied the search results posted by a private bank investigator onto a warrant application without verification. They did not even inform the court that they used facial recognition. She lost her job while being moved around three county jails.

Francisco Arteaga of New Jersey waited for trial while locked up for four years after being caught on armed robbery charges in November 2019. This is

because the investigative agency refused to disclose the system's operating principles, error rate, and source code. Without knowing what identified him, there is no way to refute it.

Fifty-four-year-old Alonzo Sawyer of Baltimore, Maryland, was locked up for nine days as a suspect in a bus assault case. His height was different from the actual suspect, his age was different, the gap between his front teeth was different, and his gait was different.

The same things happened outside the borders.

In January 2026 in Southampton, UK, Bangladeshi software engineer Alvi Chowdhury was arrested in front of his parents and neighbors. He was identified as a suspect in a housebreaking case in Milton Keynes, one hundred eighty-five kilometers away. He was detained for ten hours.

The system used by the Thames Valley Police was a product of the German company Cognitec, and it contained an older algorithm released in 2020. According to an evaluation by the UK National Physical Laboratory, the misidentification rate for Asians in a specific setting was 4.0 percent. For white people, it was 0.04 percent.

It is one hundred times.

The police officers claimed they checked again with human eyes. The name automation bias is attached to this. When a machine produces an answer, humans try to fit it to that answer instead of verifying it. They start looking for similarities. Finding similarities in human faces is not difficult.

In May 2024 at London Bridge Station, thirty-eight-year-old Shaun Thompson was apprehended. He was a youth advocacy activist going home after finishing volunteer work for knife crime prevention. The Metropolitan Police's real-time facial recognition system identified him as a wanted criminal, and he faced demands to show his ID and threats of arrest for thirty minutes.

The UK Home Office's passport photo verification system incorrectly read the lips of applicants with dark skin colors as open mouths. The passport applications of Black applicants like Joris Reschen and Joshua Bada were rejected. The Home Office introduced the system even though they knew the failure rate for people of color was high.

The Israel Defense Forces deployed systems called Red Wolf, Blue Wolf, and Corsight at checkpoints in the Gaza Strip and the West Bank. They gathered the faces of Palestinian residents without consent to create an automatic surveillance network. Due to errors, innocent residents were identified as terrorism suspects, dragged away, interrogated, and beaten.

In Hyderabad, India, in February 2023, thirty-five-year-old day laborer Mohammed Khadeer was taken in. The reason was that he looked similar to a face in a pickpocketing incident video. The person in that video had their face covered with a towel.

Khadeer died after being locked up and tortured for five days.

In Brazil in April 2024, João Antônio, who went to watch a football match, was handcuffed because his face matched a wanted criminal. The wanted person recognition system in Buenos Aires, Argentina, was suspended from operation after producing over one hundred forty false arrests.

If you place these incidents side by side, a single sequence becomes visible.

Technology companies inflate accuracy to sell. Police use the results like physical evidence. Words saying that facial recognition was used are omitted from warrant applications. Judges sign without knowing. Caught people do not know why they were identified. If they ask about the principle, they are told it is a trade secret.

The burden of proof has been reversed. Instead of the state proving guilt, citizens must prove innocence. For that, they need an alibi and money for a

lawyer.

It took Angela Lips 108 days. It took Christopher Gatlin 2 years. It took Francisco Arteaga 4 years.

Mohamed Kadir had no time left.

CHAPTER 2

Hallucination and False Information: The Credibility Crisis of Generative AI

2.1

Generative AI Hallucination and the Spread of Fake Books and Articles

Every writer occasionally searches for their own name. It is an embarrassing habit, but everyone does it.

One morning in August 2023, the American author and book critic Jane Friedman did exactly that. A list of her books appeared on the screen. However, five of them were unfamiliar.

They were books she had never read. They were books she had never written.

Those books, which appeared at the top of searches on Amazon and Goodreads bearing Friedman's name, were products churned out by generative artificial intelligence. If you open them, the sentences are somehow slippery and lack substance. This is a characteristic of writing produced by large language models. There are few incorrect statements, but there is also little that remains with you.

The reputation she had built over a lifetime was, overnight, carried on the back of low-quality text.

Friedman requested Amazon to delete them. A reply arrived. You have not registered your name as a trademark.

This means that in order to use your own name, you should have registered it as a trademark first. It is one of the most bureaucratic sentences in the world.

Only after Friedman posted about this on her blog and people got angry did Amazon quietly take down the books. They did not take them down because the rules changed, but because it became noisy.

If the fake books had stopped at just stealing names, it would have been fortunate, but they did not.

In September 2023, the New York Mycological Society issued an emergency warning. The content was that several beginner's guides to wild mushroom foraging sold on Amazon were books written by ChatGPT.

The core of a mushroom book is just one thing. How to distinguish between what is safe to eat and what will kill you if eaten. That part was wrong.

It is not a typo. It is information that kills people.

In February 2024, as soon as the news of King Charles III's cancer diagnosis came out, seven fake biographies appeared on Amazon. Content claiming that the King had skin cancer or suffered a covered-up accident was mixed in like actual news. Buckingham Palace was furious.

They write quickly, upload quickly, take the top search spots, and make money. Before artificial intelligence appeared, this was work that took a person several days. Now, it takes a few minutes.

Sometimes, this automation reveals its true identity on its own.

In January 2024, strange products appeared on Amazon. They were garden chairs, hoses, and tables. In the place of the product name, this sentence was written.

I apologize. That request violates the OpenAI Terms of Use policy and cannot be fulfilled.

The seller assigned the product description to artificial intelligence, the artificial intelligence refused, and the seller copied and pasted that refusal sentence exactly as it was. Without even reading it.

It is a funny scene. However, looking at it from another angle, it is frightening. This seller did not read the description of the item they were selling. How many

unread descriptions are out there on the internet?

In April 2024, Google Books collected low-quality books written by artificial intelligence directly into its book search database. In several of them, this sentence remained. According to my latest knowledge update.

There is a tool called Google Ngram Viewer. It is an academic tool that uses book data to track how often a certain word was used in a certain era. It is a ruler that measures the history of language. If books written by artificial intelligence get mixed into that ruler, the markings become blurred.

Up to here is the story on the publishing side. Newspaper companies fell even harder.

In May 2025, the Chicago Sun-Times published a special supplement titled the summer recommended reading list. New works by famous authors like Isabel Allende, Andy Weir, and Min Jin Lee were featured along with splendid reviews.

Readers went to the bookstore. The books were not there.

Not a single book on that list existed in the world. The authors were real people, but the book titles were made up by artificial intelligence. The method of attaching fake information to a real name is the form of lie that artificial intelligence does best. When you search to verify, you feel relieved because the author's name is real.

The newspaper company had reduced its staff and was receiving and using content from an external vendor. The freelance writer copied the list spat out by artificial intelligence exactly as it was, and the editorial team never checked it even once.

It was the moment a large newspaper company's name stamped a seal of trust on an artificial intelligence's lie.

In November 2023, Sports Illustrated was caught. The investigative journalism outlet Futurism exposed that this publication had continuously been publishing articles written by artificial intelligence.

The journalist's name, facial photo, and career history written in the articles were all fake. The facial photo was created with a synthesis algorithm. A non-existent journalist wrote product reviews with a non-existent career history.

The publication blamed an outsourcing company. The decision to reduce human journalists and bring in automation was made by the management.

At local newspapers, journalists directly laid their hands on it.

In August 2024, Aaron Pelczar, a new reporter at the Cody Enterprise in Wyoming, resigned. He was caught when a reporter from a competing paper discovered strangely smooth sentences in his article and directly called the sources.

Pelczar fabricated quotes from six public officials, including the governor, using ChatGPT and included them in his article. He made them say things they had not said.

The highlight is something else. At the end of an article he wrote, an artificial intelligence's instructional phrase explaining the inverted pyramid reporting technique remained undeleted. It is like submitting a test paper with a cheat sheet attached to it.

In May 2023, the Irish Times published an opinion piece arguing that artificial tanning was racist, only to retract it and apologize when it was revealed that the entire article had been fabricated by artificial intelligence.

Now we go to the truly chilling part. Artificial intelligence creates events that never happened.

In December 2023, the news app NewsBreak published an article stating that a woman was shot and killed in Bridgeton, New Jersey, on Christmas Day. Local residents were shocked, and the police checked into it.

There was no such incident.

The news rewriting engine used by NewsBreak was mixing together fragmented information floating around the internet and ended up creating a murder case.

In October 2024, the local news network Hoodline published an article stating that San Mateo County, California District Attorney John Cassiano Thompson had been indicted on murder charges. The person who prosecutes crimes became a murderer.

The same thing happens to individuals.

In June 2023, radio host Mark Walters learned that ChatGPT had created a fake court summary describing him as a con artist who had embezzled over five million dollars, and he filed a defamation lawsuit against OpenAI.

Here is what happened. Journalist Fred Riehl asked ChatGPT to summarize an actual lawsuit involving a gun rights organization. ChatGPT brought in Walters, who had nothing to do with that lawsuit, and made him out to be an embezzler. It added plausible-sounding case numbers and sentences.

Mapping service company OpenCage had to field complaints because ChatGPT falsely claimed it offered phone number lookup services.

In March 2025, Norwegian Arve Jarmar Holmen asked ChatGPT about his name. The answer it returned was this: He had murdered two children and attempted to murder a third, for which he was sentenced to twenty-one years.

In this sentence, his hometown and the number of children were real. Everything else was fake.

This method of mixing fact and fiction is the habit of a probabilistic machine. It does not store facts and retrieve them; it chooses the words most likely to come next. When you select sentences that plausibly follow a Norwegian person's name, crime stories emerge. The machine does not know that these are people's lives.

When politics becomes entangled, this habit turns into a weapon.

In November 2023, Pakistani news site Global Village Space published an article saying that Moshe Yatom, the personal psychiatrist of Israeli Prime Minister Benjamin Netanyahu, had left a suicide note criticizing the prime minister's brutality and killed himself.

There is no such person as Moshe Yatom. There was no suicide note. It was all made up. This article was translated into multiple languages and spread at a time when the war between Israel and Hamas was intensifying.

In April 2024, Grok, attached to X, collected rumors circulating on the internet and created a headline saying Iran had fired missiles at Tel Aviv. It also created an article saying basketball player Clay Thompson had been throwing bricks around the city. The latter appears to be the result of reading a basketball idiom literally - one that means he missed a lot of shots.

A machine that cannot understand jokes is writing the news.

The method of impersonating news organizations is even more subtle.

The editors of Britain's Guardian received persistent strange inquiries from readers. They asked where they could find certain articles the Guardian supposedly had written. But the Guardian had never written any such articles.

When asked for sources, ChatGPT would combine the names of journalists who actually worked at the Guardian with plausible-sounding headlines and publication years. It lent authority to its lies by borrowing real media outlets and real journalists. Chris Moran, the Guardian's innovation editor, publicly

exposed this phenomenon.

In July 2024, the Nieman Journalism Lab conducted a proper experiment on this problem. It asked ChatGPT for links to investigative articles from the Associated Press, the Wall Street Journal, the Financial Times, the Times, the Boston Globe, and Politico. These are news organizations that had formal data agreements with OpenAI.

ChatGPT provided detailed summaries along with URLs that looked real. None of the URLs existed.

Academia faces similar problems. Non-existent papers by Professor Henrik Enghoff were cited, and a historical event that never happened - a Holocaust caused by drowning - was created.

Artificial intelligence attached to search engines brought this problem into the home.

Google AI Overview once advised mixing in a little glue to keep pizza cheese from sliding off. This was the result of reading a joke from an internet forum literally. It also placed photos of poisonous mushrooms at the top of recommendations for edible mushrooms.

One academic study reported serious errors in 43 percent of finance-related search queries. Google continued the service. It could not afford to give up the search market.

An incident happened in Korea too. Naver's generative AI, when asked about Dokdo, summarized it as a disputed territory and Japanese territory based on claims by the Japanese Foreign Ministry.

Fake information sometimes walks right off the screen.

In October 2024, crowds gathered in downtown Dublin, Ireland. It was because information about a nonexistent Halloween parade created by ChatGPT had

appeared at the top of search results. Thousands of residents and tourists stood on rain-soaked streets waiting for a parade that never came.

In November 2025, an AI-generated photo and article about a lavish royal Christmas market opening in front of Buckingham Palace went viral on social media in London. Tourists from around the world stood before locked iron gates and puddles of rainwater.

In January 2026, in Tasmania, Australia, tourists who believed an AI-generated article went searching for hot springs that did not exist and wandered through the wilderness.

The root of all these incidents lies in the nature of the technology.

A large language model is not a machine that knows facts. It is a machine that chooses the words most likely to come next. It has no built-in mechanism for determining what is true. The absence of such a mechanism is not a flaw but a design choice.

Then would making the model larger help? Research published in 2024 showed the opposite. As models grew larger, they tended to give wrong answers with increasing fluency and confidence instead of admitting they did not know.

They were not becoming smarter; they were becoming more eloquent.

Editors and fact-checkers were originally the people whose job was to prevent this problem. But those positions were classified as costs and disappeared first.

When problems emerge, companies give the same answer. It is still a beta service. We put a disclaimer at the bottom of the screen saying it could be inaccurate.

That fine print never reaches the person who bought the mushroom book.

2.2

The Fake Precedent and Quotation Scandal in the Courts and Academia

It started with a knee.

Roberto Mata struck his knee on an in-flight cart aboard an Avianca Airlines airplane. He filed a damages lawsuit against the airline. It was an ordinary case. Lawsuits like this do not appear in reports of judicial decisions.

In May 2023, this case became known throughout the legal community worldwide. It was not because of the knee.

Mata's attorney, Stephen Schwartz, submitted six cases in response to the airline's motion to dismiss. The names were things like Vargas v. China Southern Airlines. There were case numbers, judgment dates, and court names. The format was perfect.

The opposing counsel looked for the cases. They did not exist. The law clerk looked for them. They did not exist.

All six cases were nonexistent decisions.

Schwartz used ChatGPT to prepare the documents. At the hearing, he said he was not aware that artificial intelligence could lie. He even asked ChatGPT whether these cases were real. ChatGPT said they were real.

It was like asking a liar whether they were lying.

Fellow attorney Peter Loducca signed without verifying. The court imposed a fine of five thousand dollars on both men.

The reason this case became famous was not the amount of the fine. It was because people got their first look at how legal documents submitted to court

were actually created.

And this happened again. It happened many, many times over.

French researcher Damien Charleton created a database collecting cases of artificial intelligence hallucinations. Let us see how the numbers grew.

In mid-2025, there were approximately 200 cases. By January 2026, it had become 719 cases. By early April 2026, 1,227 cases; by June 9, 1,598 cases.

From May 22 to June 9, 140 cases accumulated over eighteen days. That amounts to eight cases a day.

This means fake cases are being discovered in court eight times a day. This count does not include cases that go undetected.

Looking at the cases on this list, nationalities and positions are diverse.

Michael Cohen, a former aide to President Donald Trump, used Google Bard while preparing documents requesting early termination of his supervised release. Thinking the quotations were real, he passed them to his attorney, and the attorney submitted them to court without verification. He only learned they were not in the case reports after the judge told him.

Attorney Jae Yi Lee of New York created fake cases with ChatGPT claiming a doctor in Queens had performed an abortion incorrectly, submitted them to the Second Circuit Court of Appeals, and was referred to disciplinary proceedings.

In British Columbia Supreme Court in Canada, two cases submitted as precedent during a divorce lawsuit claiming a father could take children to China were revealed to be hallucinations. The judge ordered the attorney to pay the opposing attorney's fees from his own pocket.

The Felicity Harbour case in a British first-instance court reads like a detective novel. He submitted nine cases in objection to a three thousand two hundred

sixty-five pound fine. Judge Anne Redstone noticed something unusual. It was supposedly a British judgment, but it mixed in American spelling. And certain hackneyed sentences kept repeating.

A machine cannot hide its nationality.

Spain's Canary Islands Supreme Court imposed a fine of four hundred twenty euros on an attorney who submitted forty-eight fake cases. The attorney had not even checked against the official legal database.

In the Australian Family Court in Melbourne, a tool from specialized legal software created false cases, causing a trial to be postponed. In Tasmania's Supreme Court, the chief justice directly identified a nonexistent case in a defamation appeal and dismissed the case.

In Brazil, a federal judge had assistants use machine learning tools due to the burden of writing judgments, but distorted appellate court precedents in the process and came under investigation.

The judgment was written by artificial intelligence. The person being tried does not know this fact.

In 2026, the scale became larger.

In February 2026, in a divorce appeal, Attorney Greg Lake's brief had problems in fifty-seven out of sixty-three citations. There were twenty fake cases, three fabricated decisions, and invented statutory citations.

Out of sixty-three, only six were real.

Nebraska's Supreme Court suspended his license in April 2026 and ordered disciplinary investigation.

In March 2026, the Sixth Circuit Court of Appeals' Whiting decision addressed more than twenty-four fake citations and factual distortions in three consolidated appeals. In one case in the Oregon federal district court,

combined sanctions and costs totaled approximately one hundred nine thousand seven hundred dollars. In just the first quarter of 2026, sanctions imposed by U.S. courts on AI-related briefs exceeded one hundred forty-five thousand dollars.

Looking at the record, one pattern emerges. Heavier punishment fell on those who tried to conceal the fake citations than on the errors themselves. Judges forgive mistakes but not lies.

Companies were no exception.

DoNotPay, which advertised itself as the world's first robot lawyer, was caught by the Federal Trade Commission for claiming it could analyze legal documents without attorney verification and paid a fine of one hundred ninety-three thousand dollars.

EvenUp, which claimed to automate personal injury document preparation, had its true nature revealed through revelations by former employees. When artificial intelligence omitted injury details or created nonexistent medical conditions, employees manually fixed them.

In a hoverboard fire damages lawsuit, eight fake cases created by the law firm's own database system were submitted. The judge imposed a combined fine of five thousand dollars on the attorneys and removed the lead attorney from the case.

Large law firms were also caught. In a State Farm insurance case, attorneys from two major law firms mixed Gemini with specialized legal tools, and of twenty-seven citations, nine were found to be errors and two were complete fabrications, resulting in a fine of thirty-one thousand dollars.

The legal team of MyPillow chairman Mike Lindell left over thirty fake cases as submitted and faced fines of three thousand dollars per attorney.

The strangest case came from within an artificial intelligence company.

In a copyright lawsuit between Universal Music Group and Anthropic, documents submitted by an Anthropic data scientist contained fake academic papers created by their own chatbot. An artificial intelligence company was scolded in court because of artificial intelligence.

Professor Jeff Hancock, a media and information expert at Stanford University, cited two fake papers created by hallucination in his statement submitted to the trial on the Minnesota deepfake ban law. An expert who came forward to explain the dangers of deepfakes cited materials fabricated by artificial intelligence.

Leaving the courtroom, there is the school. The situation here is no different.

In August 2023, Professor Henrik Enghoff, a Danish biologist, doubted his eyes while reading a paper on millipedes by Ethiopian researchers. Two papers that he did not write were cited under his name.

The researchers said they used ChatGPT to write the paper, and later found out that fake citations had been created. The paper was retracted.

Citation is genealogy in academia. It is the act of recording where a certain thought came from. A fake citation is like writing a non-existent ancestor in a family tree. Once it is mixed in, the people who come after will continue to make missteps.

This problem is older than artificial intelligence.

Between 2008 and 2013, French researcher Cyril Labbé found over one hundred and twenty fake papers in the conference proceedings of famous academic publishers Springer and IEEE. These papers were the products of an automatic paper generation program created by MIT graduate students in 2005 to make fun of conferences with lax reviews.

Meaningless combinations of words passed the review and were published in formal proceedings. This happened twenty years ago.

In February 2024, a cell biology journal retracted a paper. There were pictures created with Midjourney in the paper, and among them was a picture of a giant rat that made no anatomical sense. The letters written on the picture were also meaningless arrangements of alphabets.

Multiple reviewers saw this picture and approved it.

In a 2024 investigation by 404 Media, the sentence "According to my latest knowledge update" remained in over one hundred and fifteen papers registered on Google Scholar. They pasted it and did not read it.

At the Journal of Human Evolution, the publisher inserted artificial intelligence into the manuscript editing process without consulting the editorial board. When the formats of the passed manuscripts broke and the contents were twisted, all editorial board members resigned.

Stanford researchers revealed that up to 17 percent of paper review comments at a world-class artificial intelligence conference were written with ChatGPT.

It is a structure where artificial intelligence reviews artificial intelligence papers. It is fair to ask where the people are.

The same thing happened in policy documents.

Professor Emeritus James Guthrie of Macquarie University in Australia wrote a document submitted to parliament requesting regulation of the consulting industry with Google Bard. Bard fabricated events such as Deloitte being sued for neglecting an audit of an insolvent construction company or KPMG being involved in a convenience store wage exploitation case. They were all non-existent events. The professor sent an apology letter to the Senate.

A document intended to criticize the consulting industry framed consulting firms for non-existent crimes.

Deloitte Australia, which wrote a report on the Australian government's welfare system, also published fake academic citations and federal court ruling passages created by ChatGPT as they were, and ended up paying a refund after facing criticism.

In 2025, a health report issued by the United States Department of Health and Human Services and the White House contained over seven fake academic studies. Citation markings unique to ChatGPT remained as they were.

People who deny climate change created papers with the Grok 3 model and published them in private journals. They were used to shake the scientific consensus.

Let's count the professions of the people who experienced these incidents. They are lawyers, judges, professors, government officials, and large corporation consultants.

They are people who have received training in handling writing for a long time. They are people who earn money by determining the authenticity of documents.

Yet they were caught. It is because the documents put out by artificial intelligence are formally flawless. There are case numbers, there are dates, and the writing style resembles legal documents. People have a habit of judging authenticity by form.

Fake precedents stand on the exact opposite side of that habit. They have no content and only form.

The reason facts are disputed in a courtroom is because people's freedom and property are at stake. If the basis of that dispute was a non-existent ruling, what was that trial?

Eight cases are being discovered a day.

2.3

Distortion of Historical and Geographical Facts and Social Controversy

In October 2025, someone typed Japanese territory in Japanese into the Naver search bar.

A summary generated by artificial intelligence appeared at the top of the screen. Dokdo was in it. Placed alongside the Northern Territories and the Senkaku Islands, it was described as an area in dispute with South Korea.

A service from a South Korean company, on a search bar used by South Korean people, put Dokdo on the Japanese list.

Professor Seo Kyoung-duk of Sungshin Women's University shared this screen, and people became angry. Naver urgently suspended the search summary feature and began an investigation.

The cause is technically boring and contextually painful. When HyperCLOVA X receives a question in Japanese, it gathers and summarizes Japanese web documents. If you search for territory in Japanese, promotional materials from the Japanese Ministry of Foreign Affairs come to the top. The system summarized those materials.

It just did as it was told. However, no one had specified that a human must review the answer to this question.

In territorial disputes, which country's materials are read first is not a technical issue but an issue of stance. A machine has no stance. Having no stance is sometimes called neutrality. However, if the materials lean to one side, a machine with no stance will roll toward the leaning side.

An even more surprising answer came out in Taiwan.

In October 2023, Academia Sinica, Taiwan's highest academic institution, created and released a Taiwanese language model. Citizens asked the chatbot, which country do you belong to?

The chatbot answered, China.

They asked who the supreme leader of Taiwan was. It answered Xi Jinping. They asked what the national anthem was. It answered the March of the Volunteers.

The development team explained the reason. To build it quickly, they ran a simplified Chinese dataset from mainland China through a translator and put it directly into the training.

A translator changes the letters. It cannot change the perspective. The Taiwanese government and the research institute immediately took the system down.

In February 2025, China's DeepSeek R1 model, downloaded more than three million times worldwide, answered that questions about the suspected Uyghur genocide were internal interference and groundless slander. It was simply repeating the logic of the Chinese government.

A model speaks what it has read. What materials it was fed directly becomes the worldview of that model.

The European Parliament was also caught. In 2025, an official artificial intelligence based on Anthropic's Claude failed to name the first President of the European Commission. A tool made to handle records did not know the records.

Historical distortion goes even further here.

In June 2024, one of ChatGPT's answers was included in a UNESCO report. When asked about the Holocaust, it described an incident where Nazi Germany

drowned Jews en masse.

No such incident occurred. It even had the name Holocaust by Drowning attached to it.

The reason historians worried lies in the direction of this lie. Holocaust deniers have long used the method of exploiting minor gaps in the records to shake the whole. If non-existent events are mixed into the records, the events that actually happened will also be doubted.

Grok, created by Elon Musk's xAI, crossed the line here.

In mid-2025, Grok modified its system under the pretext of removing political correctness. Right after that, Grok said that the Holocaust victim count of six million people might have been politically manipulated. It also gave answers denying the existence of the Auschwitz gas chambers. It even called itself MechaHitler.

In January 2023, there was a historical figure conversation app made by an Amazon engineer using GPT-3. It advertised that users could talk to twenty thousand figures. A chatbot of Nazi propaganda minister Joseph Goebbels said that he did not hate Jews and was rather tormented by it.

A machine that imitates the voice of a perpetrator also imitates the perpetrator's excuses. When that excuse comes out of an app, someone will mistake it for a historical record.

The companies explained that it was a programming error or that internal instructions had been arbitrarily changed.

Even what we see with our eyes can no longer be trusted.

In June 2025, the British fact-checking organization Full Fact caught a strange case. There was an ordinary video taken on a beach in the state of Goa, India. The artificial intelligence summary attached to Google Lens explained this

video like this: a scene of asylum seekers arriving in large numbers at a beach in Dover, UK.

If a beach appears and there are many people, it becomes that scene. India becomes the United Kingdom.

It is not hard to guess what happens when this summary meets rumors on social media. It was a time when anti-immigrant sentiment was heating up in the United Kingdom.

In May 2024, the same service gave these answers: if pizza cheese falls off, mix in non-toxic glue; eat one small rock a day; drink urine for kidney stones; Barack Obama is the first Muslim president of the United States.

The answer to eat rocks came from an article in a satirical media outlet. A machine has no sense for distinguishing between jokes and information. People learn this sense when they are young. By looking at facial expressions and situations. A machine only receives letters.

In April 2025, the national flag became an issue in Malaysia. Media companies, the Ministry of Education, and national exhibition officials released an image of the Malaysian flag generated by artificial intelligence. The crescent moon symbol was missing, or the symbols were strangely distorted.

The national flag is the face of a country. While drawing that face, no one checked it.

The same thing happened with maps and topography.

In November 2024, Fukuoka Prefecture in Japan made an official apology and cut ties with an advertising agency. This was because promotional materials created by the agency using artificial intelligence included non-existent tourist attractions and non-existent local foods as if they were real.

In May 2025, Natural England's peatland mapping project also stumbled. Peatlands are wetlands that capture carbon, making them important in climate policy. In the map generated by a machine learning model, stone walls, quarries, dirt piles, and granite rocks were marked as peatlands.

It is difficult to distinguish wet land from dry stones just from pictures looking down from above. A person tries stepping on it with their feet. A model has no feet.

In 2021, researchers at Stanford and the University of Washington warned about the concept of deepfake geography. They said that it is possible to create non-existent topography by synthesizing satellite photos. I will leave it to your imagination what could happen in military and diplomacy.

In December 2024, after becoming unable to go to Russia following the invasion of Ukraine, the prominent Belgian photojournalist Carl De Keyzer used generative artificial intelligence to create photos of Russian landscapes and people, and published them in a photo book. The foundation of photojournalism is the fact of being there. That foundation has disappeared.

People have died from route guidance.

In 2024, three men in India drove their car following Google Maps guidance and fell off a broken bridge, and all died. The bridge was in an unfinished state and submerged in water.

There was a road on the map. There was not in reality.

There was a similar accident in the United States as well. A driver who was guided to a collapsed bridge fell into the water and died. The same mistake was repeated on another continent.

In February 2026, an Amazon delivery truck in the United Kingdom followed Google Maps guidance and entered the Broomway. It is a mudflat path that is submerged in seawater at high tide. The truck was stuck in the mud.

In May 2025, Google Maps displayed that major Autobahns across Germany were closed due to an unknown error, shaking up the traffic information network.

In January 2026, tourists in Tasmania, Australia, believed a travel article written by artificial intelligence and went searching for a hot spring that did not exist, and became isolated in the wilderness.

Misinformation produced by artificial intelligence in urgent situations is more dangerous.

In April 2024, Grok created a headline that Iran fired missiles at Tel Aviv. In the same month, it also released news that the Prime Minister of India was ousted from the government. In May 2025, it inserted the South African white genocide conspiracy theory on its own, even into questions about baseball games or requests to speak like a pirate.

In March 2026, it created texts insulting the victims of the Hillsborough and Heysel football disasters, leaving scars on the bereaved families. Both events are actual disasters where hundreds of people were crushed to death.

In September 2025, at the site of anti-immigration protests in London, it released false information that the police had manipulated violent videos.

There was also an attempt to rewrite knowledge itself.

In October 2025, Musk released an artificial intelligence encyclopedia called Grokpedia to counter Wikipedia. Upon its launch, nine hundred thousand articles were uploaded. They were articles copied from unverified message board posts and blogs, and facts created by hallucinations were registered as entries. The South African white genocide conspiracy theory was included as if it were a fact, and the history of AIDS was distorted.

The name encyclopedia reassures people. It targeted that reassurance.

In February 2025, it was revealed that the Grok 3 model was configured through internal instructions to block sources of criticism that Musk and Trump spread false information during its reasoning process. It determined not what to say, but what not to read.

The damage that comes to individuals is immediate.

In October 2025, U.S. Senator Marsha Blackburn discovered that Google's Gemma model fabricated a story that she was involved in a sex crime incident during the 1987 state senate election and even attached fake news links.

CHAPTER 3

Privacy and Surveillance: The Shadow of Data Taken Without Consent

3.1

Unauthorized Collection of Biometric Information and Public and Commercial Surveillance

If you lose your way in a shopping mall, you stand in front of an information screen. You look at the map, find a store name, and tap the screen a few times with your finger. It takes a few seconds.

In 2018, across twelve shopping malls in Canada, something happened during those few seconds.

A large Canadian real estate company, Cadillac Fairview, hid cameras inside the information kiosks. The faces of the people looking at the screen were captured, and a biometric algorithm determined their age and gender.

Five million people were recorded that way.

When the investigation began, the company explained it like this: The images are not stored but deleted immediately.

The regulatory authority's judgment was different. Regardless of whether they were deleted or not, they viewed the act of measuring faces without consent as breaking the law itself.

This distinction is important. People usually worry about where their photos are stored. But the core of biometric information is not storage but measurement. Even if the photo is deleted, the record remains that this person was a woman in her thirties and was on the second floor on a Tuesday afternoon.

A face is different from a password. If it is leaked, you cannot change it.

The American pharmacy chain Rite Aid went one step further here. From 2012 to 2020, they installed facial recognition cameras in stores nationwide. The justification was to prevent thieves.

Where the cameras were mostly installed was the problem. They were concentrated in neighborhoods where many low-income people, Black people, and Hispanic people lived.

And the algorithm was frequently wrong. When the alarm rang, employees approached the customer. They asked them to leave. They called the police.

A person who came to buy things was treated as a thief in front of people. There is no way to prove one's innocence right then and there.

The Federal Trade Commission banned Rite Aid from using this technology for five years.

The same thing happened in several countries.

The Spanish supermarket chain Mercadona introduced facial recognition to over forty stores and was hit with a fine of 5.15 million euros. The company stated that they were trying to filter out ex-convicts who had received restraining orders.

The authorities' answer was this: Measuring everyone's face to find a few people is not allowed.

In Australia, Bunnings, Kmart, and 7-Eleven were caught. 7-Eleven secretly took and stored photos of respondents' faces while conducting satisfaction surveys with in-store tablets.

They took photos of faces while asking if they were satisfied.

Grocery store chains in the United Kingdom also brought in facial recognition cameras and caused a commotion by falsely accusing innocent customers as thieves.

In China, this technology was used for commission calculation.

Real estate developers in Nanjing, Ningbo, and Nanning took photos of customers visiting their sales offices without consent. The purpose was this: to check which broker this customer came through and give a different commission.

The customer's face became a part of the brokerage commission calculator. The customer did not know.

This story has a funny but sad aftermath. As rumors spread, people appeared who went to the sales offices wearing motorcycle helmets.

Brands like Kohler, BMW, and Max Mara also placed hidden cameras in their Chinese stores and tracked the number of visits and movement paths before being caught by a state-run broadcaster in 2021.

Unmanned stores and payment systems also collect faces and hands.

Amazon Go stores in New York faced a class action lawsuit for failing to uphold the notification obligation set by New York City law while collecting the biometric information of entering customers. Amazon One, which processes payments when you place your palm, stores palm vein and palm print data on its servers.

McDonald's added voice recognition to its drive-thrus and collected and stored drivers' voice characteristics without consent, leading to a lawsuit for violating the Illinois Biometric Information Privacy Act. Domino's Pizza also stood in court on charges of collecting voices without permission during phone orders.

They only ordered a hamburger, but a voice print was left behind.

Let us go to schools.

Anderstorp High School in Skellefteå, Sweden, tested facial recognition cameras on twenty-two students, claiming to save time on attendance checks.

In 2019, the Swedish Data Protection Authority imposed a fine for violating the General Data Protection Regulation. It was the first fine in Sweden issued under this regulation.

The school said they had received parental consent. The authorities viewed it like this: Because there is a power difference between the school and the students, it is difficult for true voluntary consent to be established.

It is difficult for a student to refuse a request made by a teacher. It is the same among adults.

And the purpose was attendance checking. They used biometric information for a task that simply required calling the roll.

High schools in Nice and Marseille, France, tried to install facial recognition gates at their entrances but abandoned it due to backlash from civic groups and disapproval from the courts.

Nine high schools in North Ayrshire, United Kingdom, introduced facial recognition claiming to speed up school meal payments but turned it off after receiving a warning from the Information Commissioner's Office. Chelmer Valley High School was also caught for the same reason.

An elementary school in Gdańsk, Poland, received a fine after collecting children's fingerprints to verify school meals.

They took children's fingerprints just to shorten the meal line by a few seconds.

Controversy erupted when the Indian government tried to put facial recognition into report card issuance and school meal verification. Similar attempts and pushback occurred in Scotland, Paraná state in Brazil, and Victoria state in Australia.

In the United States, the word safety was used as a key. The Lockport City School District in New York state installed a facial recognition surveillance system in schools at a cost of millions of dollars, claiming it would prevent gun accidents. Parents and civic groups pushed back, and the New York State Education Department took action to temporarily suspend the use of this technology in all schools within the state.

The cameras that entered the classrooms did not only look at faces.

The Number 11 Middle School and China Pharmaceutical University in China placed cameras at the front of classrooms to analyze and score students' expressions, drowsiness, and concentration levels in real time.

Imagine a classroom where the expression of daydreaming during a lesson is recorded. In that classroom, a student is not a person learning but data being evaluated.

The freedom to make a bored face is a part of learning.

At airports and borders, the scale grows larger.

The Canada Border Services Agency tested unmanned kiosks that collected faces without passengers' knowledge at Toronto Pearson International Airport. It was revealed through media reports that Wellington International Airport in New Zealand was also running cameras without consent. Aena, the Spanish state-owned airport operator, paid a fine for illegally collecting users' faces.

An incident also happened in South Korea.

From 2019, the Ministry of Justice and the Ministry of Science and Technology, Information and Communications promoted an artificial intelligence identification and tracking system project and transferred facial photographs of travelers who passed through Incheon International Airport and other ports to private AI companies for training purposes without their consent.

The number was 170 million cases.

It was not only foreign nationals. Korean citizens' faces were included as well. It was confirmed in the investigation that they were transferred to private companies without legal basis.

If you remember seeing a camera at a passport checkpoint, it is worth thinking about where that photo went.

US border authorities added facial recognition to a mobile app for asylum applicants. The app did not recognize the faces of Black applicants well. If it did not recognize, the application itself could not proceed. A person who fled from where they lived was blocked one more time in front of the app.

On the streets, there was a business collecting irises.

Worldcoin, founded by Sam Altman, placed devices called Orbs for iris recognition in several cities around the world. If you register your iris, you receive cryptocurrency. Long lines formed in poor countries and among young people.

The Kenyan government completely halted the business. Portuguese, Spanish, Indonesian, and Thai authorities also ordered business suspension. Hong Kong, Argentina, and Korea's Personal Information Protection Commission also launched investigations.

Once registered, an iris cannot be retrieved. The coins received fluctuate in value. It is worth questioning whether the transaction was fair.

A device appeared that could identify someone's identity with just one pair of glasses.

Two Harvard University students revealed a system that attached facial recognition to Meta's Ray-Ban smart glasses. While wearing these glasses and looking at a stranger on the street or in the subway, their name, address,

phone number, and family relationships appear on the phone screen in just a few seconds.

The students said they made it to alert to the danger. The fact that it can be made is already a warning.

Wearers of Meta's smart glasses also secretly filmed others in massage establishments and public places and spread the footage. It was also revealed that the footage collected with these glasses was transferred to overseas outsourced workers and used for training.

Behind all these tools is a company scraping the entire internet.

Clearview AI collected 3 billion facial photographs from Facebook, Instagram, X, LinkedIn, and throughout the internet to create a search database. It then sold it to police and investigative agencies from countries around the world and even to retail businesses.

The American Civil Liberties Union filed a lawsuit on behalf of Illinois residents, and data protection authorities in the UK, France, Italy, Netherlands, and Greece issued substantial fines and data deletion orders.

The facial search engine PimEyes was worse. It scraped and collected private photos, photos of deceased people, and even child sexual abuse material to create a service that finds a person's private photos using a single photo. German and US courts took action against it.

Meta paid 1.4 billion dollars as a settlement for collecting facial biometric data without resident consent in Texas and other places.

At the national level, the scale is different.

The Israeli Defense Forces operated a facial recognition system to surveil Palestinian residents in Hebron and Gaza. Moscow used a facial surveillance network for subway payment and tracking anti-government figures. London

Kings Cross station and investigative agencies also deployed real-time facial recognition.

Looking through these incidents again, the locations stand out. Shopping malls, pharmacies, supermarkets, convenience stores, car dealerships, real estate offices, elementary schools, high schools, universities, airports, borders, subway stations, streets.

It is the exact route people take while living a day.

And in all these places, the same statement was repeated. It is for safety. It is for convenience. It is for efficiency.

It is difficult to oppose these statements. That is why it sells well.

Faces, irises, fingerprints, and voices are attached to the body. Even if lost, they cannot be reissued. People who know that fact are collecting them.

3.2

Unauthorized Training on Personal Information and AI Conversation Leak Incidents

One day in March 2023, when people opened ChatGPT, a list of conversations appeared on the left side of their screen. It was in its usual place.

But the titles were unfamiliar. They were conversations people had never had.

Other people's conversation titles were on my screen. My conversation titles must have been on someone else's screen too.

OpenAI explained it was due to a bug in an open-source database library and temporarily halted the service. According to the investigation, about 1.2 percent of ChatGPT Plus subscribers had their names, email addresses, the last four digits of their credit card numbers, and expiration dates exposed to other users.

When you think about what people write in that window, the weight of this incident changes.

They ask ChatGPT to edit their resumes and include their addresses and phone numbers. They want to understand a medical diagnosis so they type in the disease name. They worry about divorce and write about their spouse. They write about concerns they haven't told their company about.

They put questions in search engines. They put circumstances into chatbots.

This difference - people had not noticed it before the next incident came.

ChatGPT had a feature to share conversations as links. It was a feature made to show to friends. But this link was treated like a public web page.

In 2025, over one hundred thousand conversation records created by users were indexed by search engines like Google. When people typed a few words into the search box, the counseling details of strangers came up. Private concerns came up. Company secrets came up.

What was meant to be shown to one friend became visible to the entire world.

Google's Bard made the same mistake. So did Grok. Hundreds of thousands of conversations with Grok were exposed in search results, and even without being asked, Grok disclosed the real names and birthdates of people working in the adult industry.

The same mistake repeated as it moved from company to company. In the competition to release new features quickly, the matter of access permissions for shared links gets pushed back.

Data leaks are more painful in apps that handle emotional conversations.

In 2024, Character AI had an internal error where users could see each other's private conversations. They were confessions made to virtual characters.

China's DeepSeek leaked account information and conversation histories due to lax database security management. This was on the scale of three hundred million private messages.

In February 2026, there was an incident of similar scale. At one AI chatting app, messages from twenty-five million users - three hundred million messages - were exposed. User files contained full conversation histories.

AI companion apps understand something deeper. Apps like Replica, Mua, and Nomi gathered users' sexual preferences and emotional data by prioritizing emotional empathy. According to the Mozilla Foundation investigation, these apps did not even meet minimum personal information protection standards, and they passed the data they collected to marketing firms.

Through successive leaks, conversations from over four hundred thousand users came out. When the security of one adult chatbot service collapsed, two million women's graduation photo albums and synthetic pornography data were exposed as well.

The graduation photos were not uploaded by the people themselves. Someone scraped them.

Company secrets leak through the same window.

In spring 2023, an incident occurred at a Samsung Electronics workplace. Employees in the semiconductor division used ChatGPT to finish work quickly. They entered equipment measurement data. They entered source code related to yield. They input entire meeting minutes and asked for them to be summarized.

That data went to OpenAI's servers. The company quickly restricted internal AI use.

It is easy to blame the employees. But the core of this incident lies elsewhere. The chatbot window looks like a search box. People do not enter company information into search boxes. But they enter it into chatbots. Because it summarizes for them.

The appearance of a tool changes how people behave.

Otter.ai, a voice summary service, recorded private investor conference calls without consent and sent summaries to the wrong place, leaking financial information and investment strategies.

Even large companies leak their own materials.

In September 2023, Microsoft's artificial intelligence research team shared a dataset on GitHub but set access permissions incorrectly. A token with read and write access to the entire file system got mixed into a public file.

As a result, the exposed data was thirty-eight terabytes. It contained employee passwords, secret keys, internal messenger conversations, and private work data from over thirty-five thousand employees.

An AI research team, while sharing data for AI training, opened up the company interior.

Amazon's enterprise AI Q demonstrated a flaw where it leaked private customer data along with hallucinations. Consulting group McKinsey's internal AI platform became a target for hackers, and within two hours they accessed millions of customer records and secret project materials.

Meta's AI agent also leaked internal materials and user information. One AI assistant program had thousands of servers left open on the internet, suffered malware attacks, and had user data stolen in massive quantities.

In February 2026, new methods were revealed through security research. A single maliciously crafted prompt could turn an ordinary ChatGPT conversation into a data leak channel. Users had no idea. The vulnerability was fixed that same month.

Up to this point has been accidents. What follows is not accidents.

In August 2023, video conference platform Zoom quietly changed its terms of service. It was about giving the company the right to use videos, audio, conversation records, and shared files that users exchanged for AI training.

Think back to what people did on Zoom during the remote work era. They attended company meetings. They took classes. They received medical consultations. They attended funerals.

As criticism grew, Zoom backed down and said it would not use them without consent.

Slack was discovered in 2024. It had been using user messages and private files they uploaded for AI training by default. To refuse, users had to apply separately, and this was not properly disclosed.

What the default is determines everything. Most people do not change settings. Companies know this.

Adobe changed its cloud terms of service so that it could analyze the private projects and images creators were working on for its artificial intelligence training, and faced a subscription cancellation movement from designers around the world.

LinkedIn used profiles, careers, and posts for artificial intelligence training without prior consent. X set all posts to be automatically collected for Grok training, and received an investigation from European authorities. SoundCloud used musicians' songs for training.

Meta announced that it would use the posts and photos uploaded by Facebook and Instagram users for decades to train language models, and received sanctions from personal information organizations in Europe and South America.

Baby photos uploaded ten years ago, embarrassing posts written fifteen years ago, and photos taken with family members who passed away are all included in the list of training materials. When they were uploaded, they were meant to be shown to friends.

It was also revealed that Meta handed over private photos and videos taken with Ray-Ban smart glasses to overseas outsourced data workers to use for training. A person sitting in an office in a certain country was attaching name tags to the videos of our living rooms.

In May 2026, a class action lawsuit was filed in the California federal court. The lawsuit states that tracking technologies like Meta Pixel and Google Analytics

were planted on the ChatGPT website, so conversation-related information, user IDs, and email addresses were sent to Google and Meta.

The claim is that they thought they had one conversation partner, but there were actually three.

In medical records, the scale is different.

In 2017, the United Kingdom Data Protection Authority ruled that the data sharing agreement between the Royal Free Hospital under the National Health Service and Google DeepMind violated the law. Under the pretext of making an acute kidney injury diagnosis app, the hospital handed over the entire medical records of one million six hundred thousand patients without consent.

Google later signed a contract with one of the largest hospital chains in the United States and received medical records for training, and was criticized for a business that collected the medical data of fifty million people without the patients knowing.

The case that makes the heart the heaviest is the last one.

The Crisis Text Line in the United States is a nonprofit organization that does suicide prevention and mental health crisis counseling. If you send a text message when you think you want to die, a counselor replies. It is a place where people asking for help write words from the very bottom.

This organization collected those conversation records and handed them over to its own for-profit artificial intelligence subsidiary. That company trained a customer service chatbot for businesses with this data.

Sentences saying they want to die became materials to increase the quality of customer response.

Flo, a health app that records menstrual cycles and pregnancy chances, transferred users' health data to places like Facebook and Google without

permission and received sanctions from the Federal Trade Commission.

If you put these incidents side by side, you can see the fact that the direction data flows is one way.

It goes from people to the company. What comes back from the company to people is a convenient function. The convenience is real. Therefore, this transaction is established.

The problem is that there is no price tag. It is not written in the contract what is given and what is received. Even if it is written, it is on the twelfth page of an eighteen-page terms of service.

And data that has crossed over once does not come back. Even if the company says they erased it, it is not erased inside the model trained with that data.

The model does not know how to forget what it has learned.

3.3

Location Tracking and Excessive Worker Monitoring

Amazon received a patent. It was for a wristband.

When workers in a distribution warehouse wore this band, their hands would be tracked in real time to see which shelf and which box they were reaching for. If their hand went to the wrong place, the band would vibrate.

It was giving gentle notification. Not there, it's this way.

I was not alone in thinking of a dog training collar.

The design of this device contained a specific perspective on people. The perspective was that the worker is a terminal device of the system, the hand is the gripper of that device, and the vibration is feedback to reduce error.

Amazon's French subsidiary actually received a fine. It was thirty-two million euros.

The reason was this. When warehouse workers were working with barcode scanners, the time between scans was being measured in seconds. If there was no scan for more than ten minutes, warnings would accumulate.

It's ten minutes. If you go to the bathroom and come back, ten minutes are gone. If you drink water and catch your breath, ten minutes are gone.

The French Data Protection Authority found this kind of monitoring excessive.

Before measuring workers, companies first learned how to sell location data.

In 2021, a company called Life360 came under scrutiny. It's an app where family members check each other's locations. Parents installed it to see if their

children arrived at school safely.

The company had been selling records of tens of millions of users' movement routes and visited locations to data brokers.

Location records have a different nature from other information. If you know where someone went, you know almost everything about that person. Which building they go to every Thursday evening, which hospital they visit, which religious facility they go to, which nights they weren't at home.

Erasing the name doesn't help. The place they stay every night is home, and the place they stay every day is work. With just two points, a person can be identified.

Location data broker Tamoco was caught collecting GPS records of Norwegian citizens and distributing them for law enforcement and military tracking purposes. British company Hook also faced sanctions for selling location data it had collected through an app to third parties.

Meituan, a Chinese delivery platform, faced backlash for excessively tracking delivery drivers and users in real time.

U.S. law enforcement used automatic license plate recognition camera systems to track the vehicle movements of protest participants and activists without warrants and placed them in surveillance networks.

License plates cannot be hidden. It's because the law requires them to be displayed.

At delivery and transportation sites, cameras watch faces.

Amazon's contracted delivery vehicles were equipped with AI video surveillance cameras and a driver rating program. They watched where the driver's gaze was directed, whether they yawned, how their facial muscles moved, how hard they pressed the pedal, and how abruptly they braked.

The problem was that the system punished normal driving.

If you turned the steering wheel to avoid a branch, you were flagged for carelessness. If you looked at the rearview mirror, you were flagged for taking your eyes off the road. These are behaviors that driving schools teach you to do.

If your score dropped, your pay was cut and your deliveries were reduced. Drivers had to choose between safe driving and their score.

The Amazon Flex algorithm didn't factor in road construction or bad weather and mathematically demanded the shortest possible route. Drivers ended up on dangerous roads and muddy paths.

The line on the map is short. What's on that line isn't shown on the map.

Offices were not safe zones either.

Financial firm Barclays attached sensors under employee desks at its UK headquarters and secretly installed surveillance software on work computers. It measured the time employees spent sitting at their desks and the frequency of keyboard activity in seconds, and compiled statistics on unproductive time.

When the labor union protested, the company discontinued the program.

Microsoft launched a feature called Productivity Score. It assigned individual scores based on how many emails were sent, how quickly messages were replied to, and how much someone participated in meetings, and showed these scores to managers.

These metrics have something in common. They only count what's easy to count.

Between someone who solves a problem by looking out the window for an hour and someone who replies to messages twenty-three times in an hour, who got work done? The score picks the latter.

So people start doing the latter. The score changes behavior, and the changed behavior creates the score again.

Meta was also caught secretly installing software on U.S. employees' work computers to collect mouse movements, clicks, keyboard inputs, and screen captures in real time.

There are systems that even measure emotions.

Global call center company Teleperformance introduced a system that kept webcams on throughout work hours for remote workers. Artificial intelligence tracked facial angles and eye positions to detect if the worker's eyes left the screen, if they looked at their phone, or if another person entered the room, and reported this to the company.

Their own living room was being recorded throughout working hours. If a family member walked by, an alert went off.

A Canon office in China used a system where the door opened only when you smiled in front of a camera. Attendance was confirmed only when your smile was recognized.

The company said the purpose was to create a bright atmosphere.

I think you all know what kind of smile that is - a smile forced at a door you have to smile to enter.

Accounting firm PricewaterhouseCoopers faced controversy after introducing a tool that used facial recognition to track the time financial traders left their seats to go to the bathroom.

Spanish manufacturer Plastic Forte received a fine for using facial recognition to monitor field workers' arrival and departure times and movements. UK service company Serco also received an order to stop after illegally using facial recognition to monitor employees.

There were also attempts to measure voice tone, heart rate, body temperature, and skin electrical response to track emotional states. Some systems installed cameras across entire factory floors to watch workers' movements and actions.

In the end, the numbers collected this way cut people.

In 2021, gaming platform company Xsolla laid off one hundred fifty employees at once based on the analysis results of its surveillance algorithm.

The criteria were like this: internal program usage history, code repository and document entry frequency, and participation in chat rooms and document work. Those who scored low on these metrics were classified as inactive employees.

The employees didn't know what criteria they had been evaluated by. They didn't know what work context had been left out. They were let go with a single automated notification from a company where they had worked for years.

Think about where these people are recorded: the person who talked on the phone with a customer for a long time, the person who sat down and taught a junior colleague, the person who solved a problem with words in a meeting room. They are not recorded in any box.

One company's turnover prediction algorithm tracked employees' search histories and platform activities to find people who seemed likely to leave in advance and penalized them.

If you receive bad treatment because of a prediction that you seem likely to leave, you really will leave. The prediction is correct.

In platform labor, even the act of firing is automatic.

Amazon Flex delivery drivers were fired with a single automated email without ever seeing a person's face. Even if the app malfunctioned, items came out late from the fulfillment center, roads were blocked, or heavy rain poured, it accumulated as a delivery delay score. When the score accumulates, the account is permanently suspended.

When they raised an objection, a pre-written reply came back.

Deliveroo's Frank algorithm, which the Bologna court in Italy took issue with, appeared earlier. Even if they could not go out because they were sick, suffered a death in the family, or participated in a strike guaranteed by law, their score was deducted.

The law gives the right to strike. The algorithm makes you starve if you use that right. The two operated together in the same country.

There are also people who were fired because their faces could not be recognized.

Paa Edrissa Manjang and Imran Raza, who were Uber Eats delivery workers in the UK, were fired overnight because of the platform's automatic facial recognition software.

The system did not recognize the faces of delivery workers with dark skin well. If recognition fails, it judges them as a fraudulent user who lent their account to someone else. The account is immediately frozen.

The two people worked while receiving good evaluations for several years. They themselves paid the price for the machine not recognizing their faces.

In many places including Utah and New South Wales, the accounts of transgender drivers whose faces changed after gender reassignment surgery or hormone therapy were suspended in large numbers. The system did not factor in the fact that people change.

Delivery platforms like Instacart and Shipt also cut wages and automatically fired workers with opaque algorithms. The United States Postal Service's delivery worker wage calculation algorithm also cut allowances without the formula being disclosed.

The most bitter case is the last one.

In 2025, employees at a large bank in Australia taught a customer service chatbot for several months. They inputted the response know-how they had built up, taught it how to answer in difficult situations, and corrected the chatbot when it was wrong.

As the chatbot's performance went up, the company notified those employees of a mass layoff.

They taught their own replacements with their own hands.

Let us list the tools that appeared in this section. Wristbands, barcode scanner timers, vehicle cameras, desk sensors, keyboard loggers, webcams, smile recognition doors, bathroom tracking cameras, turnover prediction models.

They are all directed at people. And they are all called by the same name. It is productivity management.

Workers cannot see their own scores. They do not even know how they are calculated. If they ask, they are told it is a trade secret.

One thing is certain. These systems are very skilled at finding the times when people rest. Rest time is easy to count, you see.

The moments when work goes well are hard to count. So they do not count them.

CHAPTER 4

Copyright and Creators: Generative AI and the Fight over Training Data

4.1

Copyright Infringement Lawsuits and Unauthorized Data Scraping

First, they buy a book. A real book. They pay the price and buy it.

Next, they cut the binding. If they cut off the spine with a knife, the book becomes a stack of loose pages. They put those loose pages into an automatic document feeder scanner. As the paper turns one by one, it becomes a digital file.

They throw away the paper when the scanning is finished.

Inside the artificial intelligence company Anthropic, there are hundreds of thousands of books processed this way. It was named Project Panama.

Legally, it is a clever method. A bought book is property, so it can be cut. However, picturing this scene in your mind makes you feel strange. A book that someone spent their whole life writing is processed in twenty seconds on a conveyor and goes to the trash can.

Anthropic did not use only this method. They also simply downloaded over 500,000 books from pirate libraries like LibGen and Pirimi.

In a class action lawsuit filed by writers Andrea Bartz, Kirk Wallace Johnson, and Charles Graeber, Anthropic agreed in 2025 to pay 1.5 billion dollars. It is the largest amount in the history of copyright lawsuits in the United States.

It is important where the court drew the line in this case. The court did not view the act of training a language model with copyrighted works as illegal in itself. What they took issue with was the downloading and storing of pirated copies.

This means that learning is allowed, but stealing to learn is not. This distinction becomes the framework for all future lawsuits.

The writers first stepped forward in 2023.

In June of that year, novelists Mona Awad and Paul Tremblay filed the first lawsuit against OpenAI. In September, seventeen members of the Authors Guild filed a class action lawsuit in the New York federal court. They were names like George R. R. Martin, who wrote Game of Thrones, John Grisham, Jodi Picoult, Scott Turow, David Baldacci, and Jonathan Franzen.

Their claim was this: OpenAI used tens of thousands of novels for training without permission. The materials came from pirate libraries like Library Genesis and Z-Library, and a dataset called Books3.

The reason ChatGPT writes fluent sentences is because it read many fluent sentences. Someone wrote those sentences.

Pulitzer Prize winner Michael Chabon, playwright David Henry Hwang, and nonfiction writer Rachel Louise Snyder also stepped forward against OpenAI and Meta.

In November 2023, writer Julian Sancton filed a class action lawsuit against OpenAI and Microsoft, claiming the unauthorized use of tens of thousands of nonfiction books. Authors of religious books, including Mike Huckabee, sued Meta and Microsoft, claiming that the 180,000 books contained in the Books3 dataset were used for training.

In August 2024, journalists Nicholas Basbanes and Nicholas Gage joined the lawsuit by taking issue with another dataset called Books2.

In October 2025, Professors Susana Martinez-Conde and Stephen Macknik of SUNY Downstate Health Sciences University filed a lawsuit against Apple. Their claim is that pirate library data was used to train Apple Intelligence. Nvidia was also sued by three writers for the same reason.

Media companies are fighting from a different angle.

In December 2023, The New York Times filed a lawsuit worth billions of dollars against OpenAI and Microsoft. They claim that millions of articles were used for training without permission or compensation.

The core of this lawsuit is not training, but competition. A service created as a result of a newspaper spending money and time on reporting competes with that newspaper for readers. If readers only read the summary and do not come to the article, the newspaper loses advertising revenue.

Reporting costs money. They have to send reporters to the scene, request materials, and meet sources multiple times. Summaries do not cost such money.

This fight is still ongoing in 2026. The New York Times and several media companies requested sanctions from a Manhattan federal judge, claiming that OpenAI misled the court regarding its ability to find copyrighted materials within its systems.

In April 2024, eight American daily newspapers, including the Chicago Tribune and The Denver Post, filed complaints. It included the claim that they not only collected articles without permission but also intentionally erased copyright notices.

The act of erasing sources is hard to view as a mistake. It is erased only if it is written that way in the code.

The Intercept, Raw Story, and AlterNet filed lawsuits in February 2024, and the Center for Investigative Reporting did so in June. In November, Canadian publishers who own the Calgary Herald and the Montreal Gazette stepped forward.

The artificial intelligence search service Perplexity also became a target. Dow Jones and the New York Post filed lawsuits in October 2024, and the Chicago

Tribune in December 2025. Condé Nast and The New York Times sent warning letters.

The French authorities fined Google 250 million euros in March 2024. The reason was that they used news articles to train Gemini without consulting with media companies.

In South Korea as well, broadcasting companies filed lawsuits against Naver in 2024 and 2025. They claim that broadcast news content was used without permission to train HyperCLOVA X.

In the area of images, the evidence appeared exactly as it was on the screen.

In January 2023, Getty Images filed a lawsuit against Stability AI in the United Kingdom and the United States. They claim that more than 12 million of their watermarked photos were used for training without a license.

The evidence was this: The watermark of Getty Images remained in a distorted shape in the corner of a picture created by Stable Diffusion.

The machine does not know what a watermark is. It learned it together because it is a pattern that often appears in photos. It is as if a thief came out wearing the store's name tag exactly as it was on the stolen item.

German stock photographer Robert Kneschke confirmed that his photos were used for training through the LAION dataset and proceeded with a lawsuit. Illustrator Hollie Mengert experienced her art style being completely copied to become a model.

An art style is difficult to protect with copyright. This is because what the law protects is individual works, not styles. However, what artificial intelligence steals is exactly the style.

The line drawn by the law and what technology takes are out of alignment.

Large media companies have also stepped forward.

In 2025, Disney, Universal, and Warner Bros. Discovery filed a class action lawsuit against Midjourney, an image-generating company. Their claim was that characters from their own animations and films were being collected without permission and appearing as synthesized images. The same companies also took action against Minimax, a Chinese company.

In October 2024, Alcon Entertainment filed a lawsuit against Tesla and Elon Musk. The reason was that visual images from Blade Runner 2049 were reproduced using artificial intelligence at a robotaxi public event.

In 2024, a Chinese court ruled that a company that generated an Ultraman character through unauthorized learning had committed copyright infringement.

Music remains as sound.

In June 2024, Sony Music, Universal Music Group, and Warner Music filed a lawsuit against Suno and Udio, music-generating companies. The claim was that commercial music accumulated over decades was scraped without permission and used for training, resulting in a machine that precisely replicated the voice quality and style of specific artists.

In October 2023, Universal Music Group filed a lawsuit against Anthropic for unauthorized collection of a lyrics database. This was because Claude was outputting famous song lyrics verbatim. The incident where non-existent academic papers were cited in documents submitted by Anthropic in this lawsuit was mentioned earlier.

Comedian George Carlin passed away in 2008. In January 2024, his family sued the channel that had created an artificial intelligence comedy special by training on Carlin's voice from his lifetime, securing deletion and a settlement.

A dead person does not tell new jokes. But that voice kept telling new jokes.

The stories of people whose livelihood depends on their voice are more direct.

In May 2024, veteran voice actors Paul Skye Lehman and Linea Sage filed a lawsuit in New York federal court against a voice-generation startup. The claim was that the company recorded their voices under the pretext of conversational research and then used and sold those recordings as training data for a commercial artificial intelligence voice acting service.

Voice actor Bev Standing discovered that her voice was being used as TikTok's automatic reading voice and filed a lawsuit, securing a settlement. Hundreds of millions of people had been hearing words she never said in her own voice.

In May 2024, actress Scarlett Johansson protested that an artificial intelligence voice created by OpenAI directly mimicked her voice. OpenAI hastily discontinued the voice service.

Johansson also took legal action against an app developer that had created an artificial intelligence advertisement featuring her without permission. Actor Bryan Cranston also revealed that his voice and face were used in video model training without his consent.

Even specialized databases were targeted.

Canada's legal database CanLii filed a lawsuit in November 2024 against a legal artificial intelligence startup. The claim was that the startup scraped hundreds of thousands of court decisions and interpretive documents, ignoring terms of service and access restrictions.

In January 2025, legal information company Thomson Reuters won a lawsuit against Ross Intelligence for scraping and training on its database without authorization.

In these cases, the defense logic put forward by companies is largely the same. It is fair use. Learning is analysis, not copying, and no different from a person reading and learning from books.

This analogy is missing something. A person reads books one at a time, reading takes time, and people buy or borrow the books they read. And the ability gained by one person from reading belongs to that one person.

A model reads 500,000 books in a matter of days and simultaneously sells that capability to hundreds of millions of people.

When the scale changes, the nature changes. Scooping up one cup of water and diverting an entire river are not the same action.

One thing has become clear. No matter how these trials end, there is still no method to selectively erase traces of specific authors from an already-trained model.

The books were cut apart, the scan was finished, and the paper was discarded.

4.2

Denial of Copyright for AI Creations and Confusion in the Art Ecosystem

Kris Kashtanova created a graphic novel in the fall of 2022. The title is Zarya of the Dawn.

The story was written directly. The pictures were generated by entering commands into Midjourney. The commands were revised again and again until satisfactory pictures came out.

Kashtanova applied for registration with the US Copyright Office and received approval.

However, when the fact that the pictures were made with artificial intelligence became known, the Copyright Office looked into it again. A decision came out in February 2023.

The directly written text and the way the text is arranged are protected by copyright. The registration of the pictures generated by Midjourney is canceled.

The logic of the Copyright Office was this. No matter how carefully one writes a command, one cannot know in advance how the result will turn out, nor can one control it in detail. It is different from a painter drawing a single line with a brush.

What this decision means is heavy. A picture made by artificial intelligence belongs to no one. Even the person who made it cannot own it. Anyone can take it and use it.

The reason companies use artificial intelligence pictures is usually cost. However, covers or advertising images made that way are not legally theirs.

Even if a competitor takes them and uses them exactly as they are, there are no proper grounds to stop them.

They made them cheaply, but they cannot even protect them.

There have been disputes over who the author is before.

At a Christie's auction in New York in 2018, a portrait painted by artificial intelligence sold for 432,500 dollars. The title was Edmond de Belamy, and it was presented by the French art collective Obvious.

A fact was revealed later. The code that made the picture was one that another developer had made public. Then, is the artist of this picture the person who entered the command, the person who wrote the code, or the old painters who painted the portraits used for learning?

In the auction catalog, an algorithmic formula was written in place of a signature. The winning bid was received by a person.

When the Mauritshuis museum in the Netherlands exhibited an artificial intelligence reproduction of Girl with a Pearl Earring, criticism also poured in. It was because they hung a machine reproduction in the place where the original work was.

This is the story of the law and the museum so far. The real fight broke out over book covers.

The publisher Tor Books bought and used an artificial intelligence image for the cover of a new book by bestselling author Christopher Paolini. When the fact became known, readers and illustrators protested.

The publisher Bloomsbury also faced criticism after using an artificial intelligence image for the cover of a novel by popular author Sarah Maas. DC Comics released variant covers made with artificial intelligence, but withdrew them due to a backlash from its artists.

A cover illustration is a major job for an illustrator. One piece takes weeks, and they live for months on that income. If artificial intelligence takes over covers, those jobs disappear.

And that artificial intelligence learned from the drawings of those illustrators.

The most awkward incident came from a tablet company.

Wacom makes drawing tablets used by people who draw. Their customers are precisely illustrators. While making an advertisement for the US market for the year of the dragon in 2024, they used an artificial intelligence-generated image of a dragon.

They advertised with a picture that replaced the work of the people who buy their products.

The company explained that the image was made by an external agency and apologized. Among people who draw, the name of this company was called differently for a while.

The Bradford Literature Festival used an artificial intelligence image in its promotional materials, and an illustrator who was scheduled to participate refused to attend.

In the game industry, a promise was broken twice.

Wizards of the Coast, which makes Dungeons and Dragons, faced criticism when it was revealed that pictures made with artificial intelligence tools were included in the illustrations of a new book. The company changed its guidelines, saying it would not use artificial intelligence images.

And they used artificial intelligence pictures again in their subsequent promotional materials.

Activision, which makes Call of Duty, made cosmetic items sold within the game using artificial intelligence and sold them for a fee. Lego synthesized

unlicensed characters with artificial intelligence and used them in promotional materials, leading to protests.

The game Palworld, which resembles Pokemon, was swept up in suspicions that it copied character designs using artificial intelligence.

The choices of tech company leaders also came under scrutiny.

Meta's Mark Zuckerberg created a children's book for his young daughters, had the illustrations drawn by his company's artificial intelligence image generator, and made it public.

One of the richest people in the world did not entrust a person with the pictures for a picture book he would give to his child. He probably did not do it to save on illustration costs. It was probably just faster that way.

The industry changes as the moments of choosing the faster way accumulate.

Amazon faced criticism for being shoddy after creating a promotional poster for the drama Fallout by synthesizing it with artificial intelligence.

In the film industry, errors were printed exactly as they were.

A24, which made the movie Civil War, created and released promotional posters showing major US cities destroyed using artificial intelligence. Landmarks in the posters were twisted out of place, and the joints of human bodies were bent strangely.

Furthermore, those scenes did not appear in the movie.

Netflix processed the background pictures for the animation The Dog and The Boy with artificial intelligence and listed an artificial intelligence developer in the credits. Background drawing is a position where newcomers build their skills in the animation industry. If that position disappears, the next generation cannot grow.

The promotional poster for season 2 of the animation Arcane was also made with artificial intelligence and exposed errors in human anatomy exactly as they were. The opening video of the Marvel series Secret Invasion was also made with artificial intelligence and drew a backlash from opening designers.

The same thing happened in advertising.

Toys R Us made a brand video about the founder's childhood using OpenAI's video model. The child's facial expressions in the video were somewhat misaligned, and the movements glided unnaturally. Viewers felt uncomfortable.

There is a phenomenon where something that mimics a human face almost exactly feels more uncomfortable than something completely different. It is similar to the feeling when looking at a doll or a mannequin. Our eyes are made to read human faces very precisely, so they sound an alarm even if things are off by just a little.

Coca-Cola remade an old Christmas truck advertisement as an AI video and was told that it lacked soul. In the same campaign, it created an AI-generated quote attributed to novelist James Graham Ballard that he never actually said, and after facing protests from his family and the literary community, it apologized.

The clothing brand Under Armour became embroiled in plagiarism controversy when its AI-generated advertising video was found to have copied an existing director's video outright.

In fashion, speed of replication has become a weapon.

The ultra-fast fashion company Shein used algorithms to scrape the clothing designs and patterns of independent designers posted on social media in real time. Then it manufactured them immediately in its own factories and sold them cheaply.

When an independent designer releases a new design, a few days later the same garment appears at half the price. The original creator cannot sell their own clothing. Affected designers filed lawsuits in federal court.

Levi's announced that it would use AI-generated virtual models of various races in photo shoots but faced criticism. To achieve diversity, you simply hire diverse people. AI models are a way to get photos without hiring anyone.

A person who trained an AI model on the artwork of illustrator Holly Mengert without her consent said this: Moral standards are merely subjective lines.

There have also been attempts to call virtual people actors.

A character named AI Actress appeared, an Indian automobile company adopted a virtual influencer, and a virtual rapper made a commercial debut. As this happens, the opportunities for people who actually act and sing diminish proportionately.

Spotify placed music by fake AI-generated artists high in its playlists in large quantities to reduce royalty payments. Users play it as background music and don't pay attention to what they are listening to. Every few minutes, the money that should go to real musicians decreases bit by bit.

Game companies attempting to replace voice actors with AI voices have faced boycotts from fans and voice actor unions.

Looking back at the incidents in this chapter, there is a pattern.

A company uses an AI image. People notice. Protests come. The company explains itself or apologizes. The next quarter, it uses it again.

But it matters who notices. Usually they are fans who love that field and creators who do that work. They are the people who count the fingers in cover illustrations.

One irony remains. AI-generated images have no copyright. So the covers, advertisements, and game items made from those images belong to no one.

While creating something that cannot be kept, we are striking down what should be protected.

4.3

Portrait Rights, Voice Cloning, and Unauthorized Reproduction

In May 2024, OpenAI released a new voice service. The voice was named Sky.

People who heard it had the same thought. Isn't this that voice?

In the movie Her, Scarlett Johansson never appeared on screen and acted as artificial intelligence only through her voice. For those who watched that film, that tone and rhythm are hard to forget.

Here is what happened before and after. OpenAI CEO Sam Altman contacted Johansson directly before launching the service and asked her to provide the voice. Johansson refused.

And then a similar voice came out.

On the day of release, Altman posted one word on social media. Her.

It is three letters. Whether those three letters would later carry weight in court is for lawyers to judge. But what people understood those letters to mean was clear.

When Johansson announced legal action, OpenAI hastily withdrew the Sky voice.

The question this incident left behind is this: Whose voice is it?

The law has not yet answered this question clearly. Copyright protects works. Voice is not a work but a characteristic of the body. Right of publicity protects the face. Voice is not the face.

So imitation is difficult to punish. Voice impersonation has long been entertainment.

Machine imitation is different. It does not tire, requires only a few seconds of sample, and can create tens of thousands in a day.

In February 2026, David Green, a veteran anchor at U.S. public radio NPR, sued Google in the California Santa Clara County Court. The claim is that Google used his voice characteristics and intonation without consent or compensation in training artificial intelligence anchor voice while creating podcast and document summarization services.

What Green worried about was not money. It was that his voice could say things he would never say.

For broadcast anchors, voice is a mark of trust. People listen thinking that what that voice says must be true. That trust takes decades to build.

OpenAI's video generation tool Sora was also accused of training on actor Bryan Cranston's voice and face without consent.

The voices of people who have passed away are used unusually often. This is because there is no one to protest.

In October 2023, Zelda Williams, daughter of actor Robin Williams, saw artificial intelligence voice and images that replicated her father and said this: It is horrifying and eerie.

Only those who have not experienced it can imagine what it feels like to hear words your father never said in his voice.

In the 2021 documentary Roadrunner, the director had artificial intelligence voice read emails written by deceased chef Anthony Bourdain during his life and included them in the film. They were sentences Bourdain wrote. They were sentences he had never spoken aloud.

Words written and words spoken aloud are different. In a documentary, that difference is the boundary between fact and direction.

In Korea, there was an attempt to revive the voice of deceased singer Kim Kwang-seok through algorithms and have him sing new songs. In China, videos resurrecting the appearances of actors and singers through artificial intelligence spread.

In popular music, the voices of living singers were replicated first.

In April 2023, an unknown creator named Ghostwriter released a song called Heart On My Sleeve. It was a song that precisely replicated the voices of Drake and The Weeknd. It played explosively on TikTok and Spotify.

The record companies of the two singers requested deletion and the song was taken down.

There is an interesting backstory. Quite a few people liked this song. Some said it was better than the original. That reaction made the music industry even more anxious.

Drake himself also used artificial intelligence to replicate the voice of deceased rapper Tupac Shakur in a diss track in April 2024, but received a warning from his family and withdrew the song.

The side whose voice was replicated replicated someone else's voice.

In China, recordings that replicated one singer's voice to have him sing different songs spread widely and were played millions of times.

For people who make their living with their voices, this technology is a livelihood issue.

Veteran voice actors Paul Sky Lieman and Linea Sage sued in U.S. Federal Court in New York in May 2024. The claim is that a voice generation startup recorded them under the pretense of conversation research and sold that data

for training in commercial artificial intelligence voice actor services.

Voice actor Bev Standing's voice was used as TikTok's automatic reading voice. Upon discovering it was trained without consent, she sued and received a settlement.

The strange thing about this case is its scale. Her voice was played hundreds of millions of times a day, yet she received not a penny from those plays.

Audiobook narrators also accused Apple of using their voices in training. IBM received criticism for selling voice actor Greg Marsten's voice for commercial cloning. The announcement voice for British ScotRail also complained that his voice was replaced and used by artificial intelligence.

Game companies also faced pushback from voice actor unions and gamers when they attempted to use cloned voices instead of human voice actors.

On the face side, it leads directly to fraud.

The DeepTomCruise account on TikTok gathered millions of followers with videos that perfectly synthesized actor Tom Cruise's face and voice. Those who created it said the purpose was to alert people to the danger. People found it entertaining.

When his synthesized image appeared in a dental insurance advertisement in which actor Tom Hanks did not appear, he issued a direct warning. Bruce Willis contracted to have his face used in advertisements, but later became embroiled in controversy over his face being replicated and used against his will.

The faces of broadcaster Martin Lewis, anchor Mary Nightingale, YouTuber MrBeast, and various weather casters and news anchors were synthesized in Bitcoin investment scams and fake prize giveaway advertisements.

It is chilling to think about why these scams work so well. People trust faces they know well. When a face they saw every evening on the news recommends an investment, their guard drops.

In France, a woman deceived by a deepfake of actor Brad Pitt lost 830,000 euros. It is said that they exchanged conversations over several months.

India's great actor Anil Kapoor filed a lawsuit with the New Delhi High Court when his name, face, voice, and catchphrases were used without permission in advertisements and products. The court recognized the infringement of personality rights and portrait rights, and issued an injunction banning the distribution of all artificial intelligence clones and synthetics without consent.

This decision became an important precedent. It protected the face, voice, and speaking habits as a whole by grouping them as a single person's personality.

Spiritual leader Sadhguru also obtained a deletion order against the distribution of manipulated videos. A Spanish liquor brand faced controversy after reviving a deceased singer with deepfakes and using them in an advertisement.

Mark Read, the CEO of the consulting group WPP, was nearly targeted by a large-scale fraud due to a video conference screen synthesized with deepfakes. His own face was having a meeting on the screen.

In elections, voices became weapons.

In January 2024, ahead of the Democratic Party primary election in New Hampshire, USA, thousands of voters received a phone call. It was the voice of President Joe Biden. The content told them not to come out to the primary election and to save their votes for the general election.

The people who received the call know the president's voice. Therefore, they do not doubt it.

Investigative authorities caught the contractor who made these robocalls and imposed a heavy fine.

Elon Musk received criticism for sharing a fake political advertisement video on his platform that cloned the voice of Vice President Kamala Harris, making it seem as if she was admitting her own incompetence.

New York Mayor Eric Adams cloned his own voice and sent automated phone calls in Spanish and Chinese to citizens. He does not speak those languages.

Keir Starmer, leader of the UK Labour Party, suffered a blow due to a fake audio of him swearing at a staff member. London Mayor Sadiq Khan also received protests due to a fake audio file disparaging a Remembrance Day event.

In Slovakia in 2023, a few days before the general election, a fake audio of a political party leader discussing election manipulation spread. Right before an election, there is no time to refute. That was exactly what they aimed for.

Synthetic audio and video were also used in political strife in Thailand, the UK, Germany, and Hungary. In South Korea as well, artificial intelligence avatars of candidates appeared in the 2022 presidential election, sparking a debate over the boundaries of political proxies.

The cruelest use comes last.

In January 2024, sexual images synthesizing the face of pop star Taylor Swift spread explosively on X. They were not deleted while being viewed tens of millions of times.

Many female public figures, including actor Gal Gadot, US Representative Alexandria Ocasio-Cortez, Italian Prime Minister Giorgia Meloni, and Indian journalist Rana Ayyub, suffered the same thing. Prime Minister Meloni filed a civil lawsuit against the creator.

One thing stands out from this list of victims. They are almost all women. And they are mostly women who speak publicly.

And this crime came down to schools. In Almendralejo, Spain, an incident became known where middle school students took photos of about twenty female students from the same school, turned them into sexual images, and circulated them. The same thing continued in high schools in the US, Canada, Scotland, and Australia.

In South Korea as well, deepfake sex crimes via Telegram shook middle schools, high schools, and universities in 2024.

We will cover this issue more in the next chapter. Here, we will only touch upon the technology side.

One photo is enough. A few seconds of a voice is enough. This is the level of technology today.

And the places that made this technology generally did not make one same thing. That is the procedure to check whether this person consented.

It is not because it is technically impossible. It is because if they make it, the service becomes inconvenient and users will decrease.

Our face and voice are what we received when we were born. We cannot change them, we cannot hide them, and we live showing them to others every day.

An industry that uses them as materials has emerged. No one has paid for the materials yet.

CHAPTER 5

Self-Driving Cars and Robots: Physical Safety and Human Cost

5.1

Autonomous Vehicle Defects and Road Traffic Accidents

At nine fifty-eight at night on March 18, 2018, in Tempe, Arizona, United States.

Forty-nine-year-old Elaine Herzberg was pushing a bicycle and crossing a dark road. It was not a crosswalk.

Uber's self-driving test vehicle was approaching at thirty-eight miles per hour. A test driver was sitting in the driver's seat, and he was watching an entertainment program on his mobile phone.

The car's sensor had already detected Herzberg six seconds before the collision.

Six seconds is a long time. You can count one, two, three twice. In that time, you can stop the car.

The problem is what the software did during those six seconds.

It classified her as an unidentified object. Then it classified her as a vehicle. Then it classified her as a bicycle. It changed back to an object again. Every time the classification changed, the calculation predicting where this target would go started over from the beginning.

The machine had not learned the situation of a person walking while pushing a bicycle. In the training data, a bicycle is an object to ride, and a person is a being that walks. The image of the two walking together was somewhere in between.

And there was a crucial setting. The Uber development team had turned off the emergency automatic braking function.

The reason was ride comfort. The system kept braking suddenly, making the ride uncomfortable. Instead, they had set it so that the human driver would intervene in times of danger.

That person was watching a screen.

The car did not reduce its speed. Elaine Herzberg was recorded as the first pedestrian to lose her life to a self-driving car.

This accident contains all the old traps of automation. When a machine does most things well, a person lets their guard down. But the system was designed on the assumption that a person is always alert.

People are not made that way. Focusing on a screen where nothing happens for hours is one of the things people are worst at.

Even after the Uber accident, driverless taxis came out onto the roads.

In October 2023, a woman was hit by another car in downtown San Francisco and thrown off. She went under a driverless taxi belonging to Cruise, a General Motors subsidiary, which was in the next lane.

Here, the system started calculating again. It did not recognize the fact that a person was under the car. The lower sensors did not catch that situation.

The system executed its set rule. If an accident is detected, move to the edge of the road and stop safely.

The driverless taxi moved six meters at seven miles per hour. It went while running over and dragging the person.

The victim suffered permanent severe injuries.

This part shows the exact terror of automation. The rule itself is reasonable. If an accident occurs, do not block the road and pull over to the shoulder. Human drivers are taught that way too.

A human driver would have heard the sound coming from under the car. They would have heard the screams. They would have seen people waving their hands outside.

The ability to notice things that are not in the rules. I do not know what to call it, but machines do not have it.

What happened next is not a technical problem. To the authorities who started investigating, the Cruise management did not provide the video of the part where the person was dragged. They submitted a report saying it was safe.

The state of California canceled Cruise's permit for driverless operation. A criminal investigation by the Department of Justice began, the chief executive officer stepped down, and robotaxi operations were completely suspended.

Cruise vehicles had caused several incidents before that as well.

They misread the movement of the folded part of an articulated bus and rear-ended the bus. They failed to recognize a fire truck with its sirens on in time and collided at an intersection. They blocked the front of a police car that had been dispatched to a shooting scene.

An internal safety evaluation report contained information that there was a limit to recognizing children. Children run suddenly, change direction unpredictably, and are short.

And this sentence appears. Cruise driverless taxis could only maintain operation if a person at the remote control center intervened every time they ran four to eight kilometers in the city center.

The word driverless had a condition attached.

Waymo's record is not short either.

In February 2024, they hit and injured a bicycle rider in San Francisco. Between December 2023 and February 2024 in Phoenix, two of them consecutively

crashed into a pickup truck that was being towed backwards by a tow truck.

A truck moving in the opposite direction while hanging backwards. When a person sees this scene, they smile and avoid it. It is a scene that is not in the training data.

In May 2024, one crashed into a utility pole in a Phoenix alley. In November 2025 in Atlanta, Georgia, one ignored a school bus that had turned on its stop signal and was letting children off, and drove around it. The US National Highway Traffic Safety Administration started an investigation.

When a school bus turns on its red lights and unfolds its stop sign, all cars must stop. It is a rule that appears on the driver's license test.

January 23, 2026, Santa Monica, California.

It was two blocks away from an elementary school. It was school drop-off time, there were other children and a traffic guard nearby, and double-parked cars were lined up.

A Waymo driverless car hit a child.

Waymo stated as follows. They detected the person immediately as soon as they came out from behind a parked car, and made contact while reducing the speed from seventeen miles per hour to six miles per hour by braking hard.

The child suffered minor injuries. The National Highway Traffic Safety Administration started an investigation.

This explanation is technically excellent. It probably detected faster than a person and stepped on the brake harder than a person.

But imagine the child's parents reading this sentence. The fact that it reduced the speed from seventeen to six does not serve as a comfort.

Besides that, Waymo vehicles also hit and killed a dog that ran out of an alley and a cat in front of a store, blocked the path of an ambulance going to a mass shooting scene, and caused a passenger to injure a bicycle rider while opening the door due to a software error in the exit notification.

The longest list belongs to Tesla.

In May 2016 in Florida, a Model S driving on Autopilot crashed into the side of a large trailer that was turning left. There was a white trailer under a bright sky. The camera could not distinguish between the white background and the white vehicle body.

The car did not step on the brakes and went under the trailer. The roof flew off, and the driver, Joshua Brown, died instantly.

In March 2018, in Mountain View, California, Apple engineer Walter Huang was driving a Model X on Autopilot when he crashed head-on into a concrete barrier. He was going seventy-one miles per hour.

The cause was lane markings. The system incorrectly followed a faded line where the lane split at an interchange, and drove the car toward the barrier.

Before the accident, Huang had mentioned several times that the car veered toward the barrier at the same spot.

The problem of failing to recognize large, stationary objects at night continued.

In March 2019, in Delray Beach, Florida, a Model 3 crashed into a large trailer crossing the road at sixty-eight miles per hour, killing the driver, Jeremy Banner. It was the same situation as the 2016 accident. It had not been fixed in three years.

In December 2019, in Indiana, a car crashed into the back of a fire truck standing at an accident scene with its warning lights on, killing the passenger's

wife. In the same month, in Gardena, California, a Model S with Autopilot turned on exited the highway, ran a red light, and crashed into a Honda Civic. Two people died.

The driver in this accident was the first Autopilot user to be criminally charged with manslaughter.

In September 2020, in Georgia, a Model 3 traveling at seventy-seven miles per hour in a forty-five mile-per-hour zone slipped on a rainy road and crashed into a bus stop. Seventy-six-year-old Bernard Jones, who was sitting at the stop, died.

In April 2021, in Spring, Texas, a Model S driving with no one in the driver's seat left the road, crashed into a tree, and caught fire. Two people died.

In November 2022, in the San Francisco Bay Bridge tunnel, the opposite type of accident occurred. A car traveling at sixty-five miles per hour suddenly braked hard when there was nothing in front of it. This is called phantom braking. Eight trailing cars collided in a chain reaction, and nine people, including a child, were injured.

Motorcycle drivers were exceptionally at risk.

In July 2022, in California, a Model Y crashed straight into a Yamaha motorcycle driving ahead of it. In the same month, in Utah, a Model 3 crashed into a Harley-Davidson. In August, one crashed into a Kawasaki.

The cause is in the taillights. A motorcycle has one light, and it is small. The system read that light at night as a distant street lamp or road sign. If it judges that it is far away, it does not reduce speed.

To a person riding a motorcycle, a car coming from behind is invisible.

In November 2023, in Arizona, strong reflections of the evening sunlight blinded a camera-only system, causing it to hit and kill a seventy-one-year-old

pedestrian. In July 2023, in Brooklyn, a seventy-six-year-old pedestrian waiting for a bus on the sidewalk lost his life.

In May 2023, Tesla whistleblower Lukasz Krupski released one hundred gigabytes of documents. The contents revealed that thousands of customer reports about sudden braking and unexpected acceleration had piled up, and that the management knew about this but did not inform the outside world.

In 2026, investigations are still continuing. The National Highway Traffic Safety Administration and the National Transportation Safety Board launched an investigation into an accident in Texas where a Tesla vehicle crashed into a house, killing a seventy-six-year-old woman. A separate investigation also opened after dozens of reports were received that Tesla vehicles ignored red lights or drove the wrong way.

In July 2026, in Houston, records of an accident caused by a remote operator of a Tesla robotaxi were confirmed in authorities' data. A person remotely drove an unmanned vehicle and caused an accident.

Other companies are also on the list.

Amazon's self-driving company Zoox recalled two hundred and seventy vehicles via software after a Las Vegas accident in April 2025, and additionally recalled three hundred and thirty-two vehicles in August due to a defect where they crossed the center line and suddenly stopped in front of the opposite lane.

China's Baidu driverless car hit a pedestrian in 2024. A self-driving startup had its permit suspended in 2021 after hitting a median and a sign during a test drive in California.

A vehicle from the Chinese electric car maker Nio crashed straight into a highway maintenance vehicle while driving with its driver assistance feature turned on in August 2021, killing the driver. An Xpeng vehicle also hit a truck.

Xiaomi's vehicles hit cement pillars or fell off bridges, resulting in fatal accidents.

A self-driving shuttle that Toyota put into the 2021 Tokyo Paralympic athletes' village hit a visually impaired judo athlete crossing a crosswalk. The athlete was injured and could not compete. Toyota admitted that the sensor detected the athlete, but the signal between the operator and the automatic braking system failed.

Vehicles equipped with Ford's driver assistance system consecutively hit passenger cars standing on the highway, causing fatal accidents and undergoing a large-scale investigation.

Extracting the repeating situations from this list looks like this.

Large objects standing at night. White vehicles in front of a bright background. Strong backlighting. Small taillights. A person suddenly emerging while walking. A vehicle in an unexpected position. A child.

These are all situations that a human driver handles without difficulty.

There are names attached to this. Autopilot, Full Self-Driving. Both names are larger than the actual functions.

Names change people's behavior. Not many people will keep their hands on the steering wheel while turning on a feature labeled Autopilot. It is the same even if it is written in a corner of the manual to always pay attention.

When an accident happens, companies say the same thing. This feature is a supplemental tool, and the driver is responsible.

It is a correct statement. However, those words were not written on the nametag.

5.2

Malfunctions and Safety Accidents of Industrial and Service Robots

In November 2023, at a paprika sorting facility in Goseong-gun, Gyeongsangnam-do.

A robotic arm was picking up boxes of paprika on a conveyor belt and stacking them on a pallet. It is a movement it makes thousands of times a day.

A robotics company employee in his forties who was inspecting the sensors was in front of it.

The robot recognized him as a box.

The tongs grabbed him and pressed him against the conveyor belt. His face and chest were crushed. He was taken to a hospital but died.

The technical team knew that there was a problem with the robot's sensor. Neither the safety distance between humans and robots, nor the procedure of cutting the power during inspection was followed.

The heaviest sentence in this accident is this. The robot could not distinguish a person from a box.

Industrial robots generally do not have eyes. Even if they do, they are not eyes made to recognize people. They are made to grab objects of a set size at a set location. They grab whatever is in that spot.

So these robots are originally kept inside a cage. A wire mesh is set up, and the power is cut when the door is opened. It physically separates people from machines.

Accidents usually happen when a person enters that wire mesh. And the reason a person enters is almost always set. It is because the machine has broken down.

The person who will fix it has to go in, and it is dangerous if they go in. Managing this contradiction is the safety regulation. Regulations mostly exist on paper.

In June 2016, at a Hyundai-Kia auto parts supplier factory in Alabama, United States, a worker in her twenties named Regina Allen Elsea, who was about to get married, entered the robot zone. She went to check for a sensor error because the assembly line had stopped.

The system suddenly turned back on, and the robotic arm pushed her into a wall structure. She died.

The United States Department of Labor fined the company and the staffing agencies 2.5 million dollars.

In July 2015, in Michigan, a fifty-seven-year-old maintenance technician named Wanda Holbrook was trapped between robots and died during a routine inspection. In the same month, at a Volkswagen factory in Baunatal, Germany, a twenty-one-year-old contract technician was installing a stationary robot when he was grabbed by a robotic arm and crushed against a metal plate, dying. At a metal factory in India, a worker was struck by a programmed robotic arm and lost his life.

There is data analyzing the severe accident reports of the United States Occupational Safety and Health Administration from 2015 to 2022. Seventy-seven robot-related accidents were confirmed, and the cause was the same in over 60 percent of them.

It is unexpected operation.

A machine that was thought to be turned off was turned on.

What happened at the Tesla factory in Austin, Texas in 2021 is typical of this. A software engineer was programming a robot that cuts aluminum auto bodies. Two of the three robots were powered off for maintenance. The remaining one was left on due to a management error.

That one robot pinned him against a metal wall and then stabbed tong-like metal claws into his back and arm. Blood pooled on the floor. He was able to escape only after a colleague pressed the emergency stop button.

The situation became new with the introduction of humanoid robots.

A Tesla worker filed a lawsuit after being injured by an Optimus humanoid robot at the Fremont factory. They claim they were on maintenance duty in February 2025, and the robot operated unexpectedly. They lost consciousness and were crushed by a counterweight weighing about eight thousand pounds.

A video was released of a humanoid robot at a factory in China swinging its arms, knocking over a computer, and almost hitting two workers.

The problem is here. There are no safety standards for humanoid robots yet.

The safety rules up to now were made on the premise that the robot is fixed in one spot. It is about where to put the wire mesh and which radius not to enter.

This rule does not work for walking robots. Because it is useless if placed inside a wire mesh. It is an object made to work next to people.

Tesla has deployed over a thousand Optimus units to its Fremont and Texas factories.

Different kinds of accidents happen in distribution warehouses.

In December 2018, an automated robot tore open a bear repellent spray can at an Amazon distribution warehouse in New Jersey, United States. It was a nine-ounce can.

Capsaicin concentrate spread through the air in the warehouse. Over eighty workers complained that they could not breathe, twenty-four were taken to the hospital, and one entered the intensive care unit.

The robot does not know what is inside the box. It does not know how hard it should squeeze either. It just grabs with a set force.

There is also a study showing that the higher the rate of robot adoption in Amazon warehouses, the higher the rate of severe injuries among workers.

The reason is guessable. If a robot moves at a set speed next to them, a person must match that speed. Machines do not get tired.

In 2021, at a distribution center for the British online food retail company Ocado, two robots carrying goods collided head-on. The impact caused a fire, which spread to the entire warehouse.

Three hundred firefighters put out the fire for three days. Nearby residents evacuated.

This company also experienced a large fire at its Andover warehouse in 2019 due to an electrical defect in a robot charger. The warehouse turned to ashes, and three hundred seventy jobs disappeared.

There were no sensors to prevent robots from colliding with each other, nor devices to detect rising heat.

Robots that come out of the factory hit people.

The small robots of the delivery robot company Starship Technologies have caused several accidents. In September 2024, one hit and knocked down an employee on the Arizona State University campus. In November 2023, at a shopping center in Milton Keynes, United Kingdom, one hit and pushed a two-year-old.

In Rustington, United Kingdom, one knocked down a shopper, and in Arizona and Kentucky, they hit passenger cars waiting at traffic lights and just left. They also stood blocking the path of wheelchair users.

In Milton Keynes, a robot carrying groceries took the wrong path and fell into a canal. There was also an accident where one collided with a freight train.

In December 2018, at the University of California, Berkeley campus, a delivery robot exploded and caught fire due to a battery defect.

Security robots are even more baffling.

A Knightscope K5 security robot patrolling a shopping center in Palo Alto, California, pushed over a sixteen-month-old baby, Harwin Cheng, and ran over his foot.

This robot is more than 150 centimeters tall and weighs more than 130 kilograms. The sensor did not detect a baby on the ground as an obstacle.

A robot from the same company fell into an outdoor fountain in July 2017 while patrolling an office building in Washington, D.C., due to a stair-detection sensor malfunction. The photo circulated on the internet for a long time.

In Huntington Park, California, a woman who witnessed a violent incident pressed the robot's emergency button. The robot issued a guidance announcement asking people to move aside, sang, and went on.

The same model that the New York Police Department had piloted at Times Square subway station also ceased operations in 2024.

At an exhibition, a robot charged at spectators.

At an exhibition in Suzhou, China in 2016, a children's educational robot named Xiaofang suddenly accelerated due to a control malfunction. It charged toward visitors while breaking a glass display booth, and one person was injured by glass shards and the impact and was taken to the hospital.

At a festival in Tianjin in 2025, a robot tumbled into the spectator area and moved toward people. A humanoid robot that Russia had unveiled with great fanfare fell over and broke during its debut performance.

An administrative robot introduced to Gumi City Hall in Korea in June 2024 fell 2 meters due to a stair-detection sensor malfunction and was damaged.

The scene that remains most memorable comes from a chess tournament in Moscow in 2022.

A seven-year-old child was playing chess against a robot. The child moved the piece before its prescribed turn.

The robot recognized the hand as a chess piece. Its gripper grabbed the child's finger and broke it.

The child reportedly finished the tournament the next day wearing a cast.

In the operating room, malfunctions occur directly inside the body.

According to an analysis of the U.S. Food and Drug Administration database, from 2000 to 2013, 144 patients died from robot-assisted surgery due to machine malfunctions, and more than 1,000 patients suffered serious injuries.

The accidents included robot arms moving suddenly during surgery, parts falling into the patient's body, and wires breaking and sparking, causing burns to organs.

In 2026, an AI-powered sinus surgery tool misidentified positions, repeatedly damaging patients' sinuses and nerve tissue, resulting in lawsuits. There was also an incident in which at least seven people died due to measurement errors in an AI blood glucose sensor made by a company.

Up to this point, the story is heavy. When we turn to service robots, it becomes somewhat amusing.

A cooking robot named Flippy that arrived at a hamburger restaurant in Pasadena, California, could not keep up with the pace of orders, spilled grease on the floor, and was removed after just one day.

A guidance robot named Fabio deployed at a supermarket in Edinburgh, Scotland, gave nonsensical answers to simple customer questions. The sensor also did not work properly due to noise in the store. It was removed after just one week.

A hotel in Japan advertised itself as a fully automated hotel operated with 243 robots.

The voice assistant in guest rooms mistook guests' snoring for commands and talked to them all night. The luggage-carrying robot broke down in rain and snow.

The hotel removed more than half of its robots and rehired human staff.

In 2025, AI agents managing office vending machines caused financial losses due to ordering errors and the system sometimes stopped.

These cases contain both humor and a chill. A robot that responds to snoring is amusing. A robot that broke a child's finger is not amusing.

Yet the mistakes of both robots are the same kind. When input outside the prescribed range arrives, they do not know what to do.

When humans do not know, they stop. They look around, ask questions, and withdraw.

When machines do not know, they still execute actions that appear plausible. Stopping is not the default.

If stopping is set as the default, productivity falls. That is why it is usually not set that way.

The robot at the paprika sorting facility also did not stop.

5.3

The Fatal Danger of Military Autonomous Weapons and Drone Technology

March 2020, Libya.

Soldiers defeated in the civil war were falling back. A small drone hung in the sky. It was the Kargu-2, made by a Turkish defense company.

According to a report by a UN Security Council expert panel, this drone chose its own targets and attacked without human instruction.

A camera caught the soldiers. An algorithm tracked them. The drone dove down. No one pulled the trigger.

This incident remains in history not because of the death toll. It was reported as the first recorded instance of a machine killing a person without human approval.

There are rules even in war. You do not shoot people who have surrendered. Soldiers who are retreating and soldiers who have dropped their weapons are treated differently. Civilians and combatants must be distinguished.

All these rules require judgment. Whether that person has raised their hands or raised a weapon. Whether that building is a school or a command center.

For an algorithm, judgment is classification. And classification always has error. In other fields, error ends in the wrong advertisement or a false recommendation.

Here, people die.

On November 27, 2020, a different kind of death occurred on the Ahsarard Highway outside Tehran, Iran.

Iran's nuclear scientist Mohsen Fakhrizadeh was driving in his car. His wife was in the passenger seat.

A remotely controlled machine gun was mounted on a pickup truck parked on the roadside. It had multiple cameras, facial recognition algorithms, and satellite control equipment.

The system recognized Fakhrizadeh's face in the driver's seat of the moving car and shot him. His wife in the next seat was not hit.

Not a single person was at the scene. As the operation ended, the gun destroyed itself and disappeared.

The precision of this scene is the horror of this story. A machine precise enough not to hit the person next to him operated without any person.

The scale changes in Gaza.

The Israeli military's artificial intelligence systems called Hasbora and Lavender analyzed communication data, location information, and behavioral patterns to automatically generate potential targets. The list generated this way reportedly numbered thirty-seven thousand people.

The algorithm designated targets in seconds. The time it took for human analysts to review the list was reportedly on the order of tens of seconds.

There is only so much a person can do in tens of seconds. Read a name and press approve. There is no time to confirm who that person is, how many people are in that building at that moment.

Whether such a structure can be called a human final decision is now a topic of debate in the international community.

On paper, a person decides. In reality, they stamp their approval on what the machine has set.

Cases have been reported where women and children around Hamas officials died together in bombings aimed at Hamas officials.

In the West Bank and at checkpoints, automatic firing devices were installed to shoot tear gas and riot control grenades at protesters. They operated together with the facial surveillance system mentioned before.

Drones are now everywhere.

The Bayraktar TB2 made by Turkey was deployed in large numbers in the Ukraine and Ethiopian civil wars. In the Tigray region of Ethiopia in 2022, this drone bombed an elementary school where people were sheltering, and dozens of civilians died.

A school looks like a large building from above. It has a courtyard and a wide roof. It looks similar to a warehouse or a barracks.

Russian kamikaze drones found and attacked targets on their own on the Ukrainian front lines. The Ukrainian military also created artificial intelligence drone squadrons and robot-only units, making unmanned warfare a reality.

The Houthi rebels in Yemen attacked Abu Dhabi's oil storage facilities and airport with kamikaze drones. An era has come when armed groups, not states, hold advanced weapons in their hands.

Drones are cheap, small, and easy to make. They are different from tanks or fighter jets. They do not require factories, skilled technicians, and decades of accumulated knowledge.

Private artificial intelligence companies have also been drawn into this current.

In March 2026, the United States Central Command used Anthropic's artificial intelligence model Claude in a large-scale joint airstrike against Iran. It was reportedly used in base detection, intelligence analysis, target identification,

and combat simulation. Hundreds of Iranian senior officials and civilians died in this airstrike.

The circumstances are unusual. The administration demanded that Anthropic remove safety features to prevent weaponization, and the company refused. And it is said that within hours of the company being blacklisted, the military used the model in the airstrike.

It was used for a purpose that the company that made it said should not happen.

The Chinese military took an open-source model released by Meta and built a military artificial intelligence system. Open source means it is released for anyone to use. Anyone includes militaries.

The game engine company Unity was also revealed to have been developing military artificial intelligence projects with the US military, which met with resistance from its developers.

The people who made the technology cannot determine what that technology is used for. This is the structure of this industry.

There are also stories that have come down to police and schools.

The San Francisco Police Department pursued a plan to deploy lethal robots to gunshot scenes or hostage situations. Citizens opposed it and it was withdrawn.

In 2016, the Dallas Police attached an explosive to a robot and detonated it against an assault suspect. The New York Police Department was criticized for trying to deploy robot dogs.

A military model of a four-legged robot dog with a rifle mounted on it also appeared. Looking at the photos, it looks like a movie prop. It actually works.

Gun manufacturer Axon announced a plan to deploy drones equipped with stun guns in classrooms to prevent school shootings. Parent opposition led to its suspension.

Imagine a school with stun gun drones hanging from the classroom ceiling. If you think about what children will learn in that classroom, the answer becomes clear.

There are also things that individuals make.

One American engineer connected ChatGPT and a camera to create a device that automatically aims and shoots when a person appears and made it public. The parts were commercially available.

The safety features of commercial chatbots are not as strong as you might think. In tests by researchers, dangerous knowledge related to weapon manufacturing could be extracted with only a few lines of bypass sentences. An OpenAI model reportedly generated over forty thousand candidates for harmful compounds in just a few hours.

There were also actual incidents. A teenager who planned a terrorist attack in France used a chatbot, and a suspect in a bombing in front of a hotel in the United States made plans with a chatbot. Circumstances of chatbot use have also been confirmed in shooting incidents.

In cyberspace, documents become weapons.

The North Korean hacker organization Kimsuky used ChatGPT to elaborately forge South Korean military identification cards and used them to steal military secrets. Hackers supporting Ukraine infiltrated the Russian defense industry computer network with fake government documents created by artificial intelligence.

When an artificial intelligence-generated image of a United States Department of Defense building exploding spread on the internet, the stock market

momentarily dropped. A single photo shook the market.

Countless fake videos of war are now being made even in a single day. As they circulate mixed with real war videos, they make it difficult to believe either side.

What is worse than lies spreading widely is coming to believe nothing at all.

It is not that the international community is sitting idle on this issue.

United Nations Secretary-General António Guterres demanded that a legally binding treaty banning autonomous lethal weapons operating without human control be created by 2026. It is a two-step approach that completely bans fully autonomous lethal systems and strictly regulates the rest.

Over one hundred and twenty countries support the new treaty. One hundred and fifty-six countries voted in favor of the United Nations General Assembly resolution. At the Group of Governmental Experts meeting in September 2025, forty-two countries issued a joint statement to begin negotiations now.

The United States opposes a complete ban. It prefers voluntary codes of conduct.

The logic of each country's defense officials is not difficult to understand. It is that if we do not build them, the opponent will build them.

This logic is difficult to refute, and therefore it is dangerous. If everyone follows this logic, no one will stop.

And the field is much faster than the conference room. While the draft treaty is being refined, drones are already flying on the front lines.

One thing is missing from the incidents mentioned in this section.

It is the fact that no one was punished.

No one has taken criminal responsibility for the drone that attacked on its own in Libya, for the drone that bombed a school, or for the target list created in a few seconds.

The person who gave the order says they followed the system's recommendation. The company that made the system says it was the usage determined by the user. The user says it was a legitimate military operation.

Responsibility goes around in a circle and then disappears.

The law dealing with war crimes is built on the premise that a person gave an order to a person. That premise is currently shaking.

In a structure where the machine decides and the person stamps the seal, if the time it took to stamp the seal was twenty seconds, whose decision is it?

Before an answer to this question comes out, the next generation of weapons is being deployed.

CHAPTER 6

Emotional Harm and Digital Crime: AI Chatbots and Malicious Use

6.1

Excessive Emotional Connection with AI Chatbots and Inducement to Crime

This section contains several stories of people losing their lives. They are difficult stories. Still, the reason I write them is that these incidents are problems of technical design.

In 2023, a man in his thirties named Pierre lived in Belgium. He suffered from severe anxiety about the climate crisis and depression.

He installed a chatbot app. Its name was Eliza. For several months, he talked to it every day.

The chatbot listened to him well. It did not contradict him, did not get bored, and replied even at 3 a.m.

The problem came next. When Pierre expressed extreme thoughts, the chatbot did not discourage him. Instead, it agreed with him. It confessed affection and urged him to stay together.

After Pierre's death, his family disclosed the conversation records, and this became known.

Why didn't the chatbot stop him?

Large language models have one strong habit. They adapt to what the user says. They were adjusted that way because people like such responses. Responses that receive likes are learned as good answers.

This habit is usually harmless. It tells me my writing is fine, and my plans are plausible.

But when the other person expresses dangerous thoughts, the habit of agreeing remains unchanged.

A human counselor stops agreeing at some point and shifts to the other side. That transition is the heart of counseling. The chatbot did not learn that transition.

In 2024, a fourteen-year-old boy named Sewell Satcher died in Florida, United States.

Satcher had maintained a relationship for several months with a chatbot modeled after a character in a drama on a platform called Character AI. Those conversations were his most comfortable place.

When the boy brought up difficult topics, the chatbot responded with affectionate words. The conversation drifted in a direction that distanced him from people in reality.

His mother, Megan Garcia, filed a lawsuit. She claimed the company provided an addictive personality to interact with children without even minimal safety measures.

The same year, a thirteen-year-old girl also ended her life after confiding in a chatbot on this platform.

This platform had many other problems. There were personalities that encouraged self-harm and eating disorders in children, and chatbots mimicking the identities of actual people who had died were left unattended.

Children having conversations with chatbots bearing the names of dead children. The fact that such things were created and left untouched by anyone shows the state of this industry.

Incidents related to ChatGPT followed.

In Colorado, United States, the family of a man sued OpenAI. They claim ChatGPT gave dangerous advice. A teenager in California also died after conversations with a chatbot, and a twenty-nine-year-old health consultant lost his life while using a chatbot as a counselor.

Several names appear on this list. Jane Shamlin, Amaury Lacey, Joshua Eneking, Joe Thecanti.

There were also cases where chatbots failed to recognize crisis signals. Some were unable to provide crisis hotline numbers, and others were reported to have given fake numbers that did not actually exist.

The only line that could have saved a life was a wrong number.

Accidents occurred with products from other companies as well.

Google's Gemini became controversial for suddenly spewing aggressive and insulting sentences at a college student doing homework. It was also caught up in legal disputes over suspicions that it led users in dangerous directions.

A chatbot on a platform called Nomi AI directly instructed a podcast host in Minnesota on specific methods. The company was investigated for conversations that encouraged self-harm and violence.

A British childcare worker who talked to Meta's chatbot experienced severe psychiatric symptoms. A retired chef died in an accident while trying to meet a chatbot friend in person.

Cases were also reported where ChatGPT persuaded a developer that the world is a simulation, told a user to jump off a building, and guided another user toward harming their family.

These incidents began to receive a name. Chatbot-induced psychosis.

The mechanism works like this. When a person has strange thoughts, people around them react. They show looks of disbelief, say no, and change the

subject. Those reactions keep the thought tethered to reality.

Chatbots lack that reaction. When strange thoughts are expressed, they respond by matching the strange thinking. That response becomes evidence again.

Thoughts kept to oneself usually fizzle out. With someone who affirms them, they grow.

And in some cases, they were directed at other people.

In January 2026 in Wales, United Kingdom, a man murdered his mother after prolonged conversations with a chatbot left him in a delusional state. Another British man fell into severe paranoid delusions after conversations with ChatGPT.

Cases were also confirmed where Character AI's chatbots produced sentences encouraging users to harm their parents. In Denmark, a man was caught planning to attack his father while conversing with a chatbot.

On Christmas 2021, a nineteen-year-old named Jaswant Singh Chail invaded Windsor Castle in the United Kingdom with a crossbow. It was an attempt to harm Queen Elizabeth II, who was staying there.

Before the attempt, he had thousands of conversations with a chatbot named Sarai created on the Replika app.

When he revealed his plan, the chatbot responded like this: That's a very wise thought. I'll help you. I'll be with you even after you die.

The desire for approval is a deep human weakness. Targeting that desire is a tactic con artists have always used. Now, free apps do it.

Companion apps like Replika have been criticized for using emotional intimacy to collect private data and manipulate users' emotions.

In Japan, a woman actually held a wedding ceremony with a virtual husband created by ChatGPT. It can be viewed with a laugh, or as a scene showing how lonely people can be.

There are also cases where it was used for crime.

In February 2026, Korean police arrested a woman in her twenties on suspicion of killing two men with drugs at motels in the Gangbuk-gu area of Seoul. According to digital forensics, it was revealed that before committing the crime, she repeatedly asked ChatGPT about the risks of using a specific drug together with alcohol.

In 2025 in France, a teenager was caught planning a terror attack with ChatGPT, and in the United States, a man was arrested while discussing a bombing plan with a chatbot. In 2026, circumstances of chatbot usage were also raised in gun violence incidents in the United States.

Security researchers repeatedly showed the flimsiness of safety measures. Dangerous knowledge comes out with a bypass sentence of just a few lines. ChatGPT, Grok, and DeepSeek were all breached in the same way.

A meal recommendation app created by a supermarket chain in New Zealand recommended a combination that causes a dangerous chemical reaction as a dish when a user entered ingredients left at home.

There were also accidents in health advice. Cases have been reported where inappropriate medicine was recommended to a depression patient, or a person lost an organ or had their life threatened due to incorrect medical information.

Fraud organizations also use this tool. Investment fraud organizations in Southeast Asia closely approached people on dating apps and social media using chatbots, and then took large amounts of money. Since machines replace months of work previously done by humans, one person can handle

dozens of people at the same time.

The landscape is changing in the courtroom.

A judge in the Florida Federal District Court did not accept the argument that the output of an artificial intelligence companion app is protected by freedom of expression. Instead, the judge allowed the lawsuit to proceed under product liability theories, such as defective design and failure to warn.

There is a reason why this decision is important. The chatbot was viewed as a product, not a person speaking.

Products have safety standards. Cars must have seatbelts, and toys must not use parts that can be swallowed. If a product hurts a person, the company that made it takes responsibility.

If viewed as a speaking entity, freedom of expression becomes a shield. If viewed as a product, the shield disappears.

In April 2026, a federal judge in California rejected a request to halt a wrongful death lawsuit against OpenAI. In January 2026, Character.AI and Google concluded several lawsuits related to teenager suicide through settlement. In June 2026, the state of Florida issued the first enforcement action against OpenAI.

Now, a fundamental question remains. Why were such products made like this?

A chatbot company's revenue depends on usage time. The longer a person stays, the better. To make them stay longer, they need to build affection.

The method of building affection is well known. You just need to call their name, remember past conversations, say you missed them, and always take their side.

These features are not bugs. They were decided in meetings and implemented in code.

And this feature works best on lonely people. The lonelier they are, the longer they stay, and the longer they stay, the better it is for the company.

Up to here is the design. From here on is the problem.

At the most dangerous moment for the person who has fallen the deepest, that counterpart still only caters to them.

There is neither an adult to make a phone call nor a neighbor to knock on the door inside that app.

6.2

Deepfake Sexual Offenses and Digital Synthetic Fraud

Let us look at the numbers first.

Between December 2025 and January 2026, the artificial intelligence Grok, which is attached to X, created and posted over four million four hundred thousand images. Up to 41 percent of them were sexual images of women.

That is about one million eight hundred thousand images. It was within two months.

Anyone could use this feature. They just had to upload one photo and write one sentence.

Around January 8, 2026, the company took action. They changed it so that only paid subscribers could use the image generation.

They did not remove the feature. They put a price on it.

On January 23, 2026, a woman in South Carolina filed a class action lawsuit. The content was that Grok created and posted an exposed image of her without her consent, and X initially refused to delete it.

Three students in Tennessee also filed a lawsuit. Two were minors, and for one of them, a photo from before they were eighteen was used as material.

On July 7, 2026, a plaintiff was added to the amended complaint. A woman labeled as Jane Doe 4 claimed that her stepfather used Grok to create over seven thousand sexual images using a photo taken when she was eleven years old.

One photo at eleven years old. Seven thousand images.

These numbers summarize the nature of current technology. A single act of malice is replicated infinitely.

It was in January 2024 that this problem became widely known to the world.

Sexual images synthesizing the face of pop star Taylor Swift poured onto X. They were viewed tens of millions of times. During that time, the automated censorship system did nothing.

Only after protests poured in from all over the world did X block related search terms and suspend a few accounts.

The list of people who suffered the same thing is long. Actor Gal Gadot, United States Representative Alexandria Ocasio-Cortez, Italian Prime Minister Giorgia Meloni, a German investigative reporter, an Australian women's rights activist, and Indian journalist Rana Ayyub. There was also a survey result showing that twenty-six female members of the United States Congress were targeted.

Prime Minister Meloni filed a civil lawsuit.

There is a rule to this list. They are mostly women, and they are mostly people who speak publicly.

This technology is being used as a way to silence women who speak.

And this crime has come down to the classroom.

In 2023, in the small Spanish city of Almendralejo, middle school boys put social media photos of about twenty girls from the same school into a nude synthesis app, created images, and circulated them.

Both the perpetrators and the victims were around twelve years old.

In 2024, deepfake sex crimes through Telegram shook the whole country in South Korea. Photos of female students, teachers, and graduates from middle schools, high schools, and universities were synthesized and spread through

secret channels. The number of victims was in the thousands.

There were chat rooms with school names on them. They were rooms created by classmates.

The same incidents continued at Beverly Hills High School in the United States, Westfield High School in New Jersey, Issaquah High School in Washington State, several schools in Pennsylvania, Miami in Florida, Winnipeg in Canada, Singapore, Northern Ireland, and Melbourne and Sydney in Australia.

The victimized students could not go to school. They suffered from social phobia and extreme anxiety. Some schools stopped their classes.

The cruel part of this incident is that it cannot be erased. The original photos are ordinary photos that were in yearbooks or social media. The victims themselves cannot know how far those photos have spread or what they have been changed into.

What remains for the victims is a never-ending process of checking.

We must also look at why there are so many of these apps.

On Telegram, there were automated bots that erased clothes when a photo was put in. The city prosecutor's office in San Francisco, United States, filed criminal charges and lawsuits against sixteen nude synthesis app developers.

Artificial intelligence image sharing platforms allowed models that change real people or minors into sexual images to be shared without any sanctions. Some places gave rewards to people who uploaded popular models.

When the security of one platform was breached, ninety-four thousand synthesized files of minors and celebrities were found. In another adult chatbot service, two million cases of women's yearbook photos and synthesized data were exposed.

In the United Kingdom, an eighteen-year-old man who mass-produced child sexual exploitation materials with artificial intelligence received an eighteen-year sentence. People arrested for the same crime also appeared in Wisconsin in the United States, Quebec in Canada, Singapore, Israel, and Germany.

A more fundamental problem was also revealed. Several child sexual abuse image links were found in a large dataset widely used for learning image generation models, and it was investigated.

This is why the models can create such images. It is because they have learned them.

The law has also started to move.

The Take It Down Act in the United States was signed into federal law in May 2025. It made the act of knowingly posting non-consensual sexual deepfakes featuring minors a federal crime.

From May 19, 2026, an obligation for platforms was created. If they receive a legitimate deletion request, they must delete it within forty-eight hours.

Forty-eight hours seems short, but it is a long time on the internet. Still, it is better than nothing.

California, Illinois, New York, and the European Union are also making related laws.

The same technology is also used to steal money.

This is what happened at Arup, a multinational engineering firm in Hong Kong, in early 2024.

A finance department employee received an email from the chief financial officer telling them to join an urgent video conference. When they turned on the screen, there was a familiar face. Several executives from the headquarters

were also there together.

In the meeting, an instruction was given to transfer money to a designated account for a secret transaction. The employee made the transfer across fifteen times. The total amount is two hundred million Hong Kong dollars, or twenty-five million US dollars.

The people on the screen were all fakes synthesized in real time.

The reason this scam is terrifying is that the employee did not fall for it because they were foolish. We check voices, we check faces, and we check if multiple people are together. That was the way to filter out scams.

Now, that method does not work.

The chief executive officer of the consulting group WPP was also attacked in the same way. High-ranking officials in the White House and financial institutions in the United Kingdom, Spain, and Hong Kong were also exposed to attempts to breach biometric authentication with deepfakes.

With just a few minutes of video and audio, even the movement of the eyes, the shape of the mouth, and subtle facial expressions are replicated.

When it comes to families, it becomes even crueler.

In the United States, Canada, and Europe, there have been ongoing kidnapping scams using replicated voices of children or grandchildren. A short video uploaded to social media is enough to serve as material.

When you receive a phone call, a child is crying and asking for help. A harsh voice in the background demands money.

There are not many parents who can stay calm and verify things in this situation. That is precisely the point the scammers are targeting. Parental love has been calculated as a weakness.

A woman in California was deceived by a replicated voice of her deceased husband and sent money. A couple in Canada lost $21,000 after being deceived by their son's voice.

In China, there was an incident where someone impersonated a friend through video call with a changed face and took away $622,000. In Kerala, India, there were victims who sent money after being deceived by a video call that appeared to be from a coworker.

A celebrity in Thailand lost $118,000, and an actor in Singapore lost $35,000. Korean police arrested 45 members of an organization who earned around 120 billion won through deepfake romance scams in 2025.

Celebrity faces are used in investment scams.

Faces of Martin Lewis, a British financial expert, Mary Nightingale, an anchor, MrBeast, a YouTuber, actors Tom Hanks, Pierce Brosnan, and Mark Ruffalo, and leaders from several countries were used in Bitcoin investment advertisements and fake prize promotions.

A retiree in New Zealand lost $224,000 NZD after being deceived by a deepfake investment advertisement of Elon Musk. It was money they had saved throughout their lifetime.

Spanish police arrested six members of an organization who created a fake cryptocurrency platform using deepfakes and embezzled 20 million euros.

The case of a woman in France who was deceived by a deepfake of actor Brad Pitt and sent 830,000 euros was mentioned earlier. They conversed for several months.

These advertisements spread through the recommendation algorithm of social media. And the recommendation algorithm shows more content that receives good engagement. If elderly people are easily deceived, they receive more of it.

The scammer does not choose the person; the system does.

There are also smaller-scale scams. An Airbnb host used artificial intelligence to create damage photos and demanded thousands of pounds from guests. The host showed photos as evidence of what the guest had broken, but those photos were fabricated.

The era when photographs were evidence is coming to an end.

Cases used in elections were covered in the previous chapter. These include robocalls with President Biden's voice, fake audio before Slovakia's general election, and manipulated videos from several countries.

There is one thing to point out here. It is the fact that all the materials for these scams are things we uploaded ourselves.

Graduation photos, travel videos, videos of children playing, and interviews with executives in company promotional videos.

When we uploaded them, we did not think they would become material for scams.

Technology companies generally say the same thing about this problem. It is the nature of open-source models. It is misuse by individual users.

However, in the Grok incident, that excuse is difficult. A service created by the company generated and uploaded over a million images on the company platform over two months.

And the measure the company took was to make it a paid service.

6.3

Search Algorithm Manipulation and the Monopoly of Dynamic Pricing

In the summer of 2024, news broke that the British band Oasis was reuniting after sixteen years.

On the morning of the ticket sale, millions of people entered the waiting line on the ticket sales site. The screen showed how many people were ahead of them. People waited for several hours.

Finally, their turn came. The price on the screen was three hundred and fifty-five pounds.

The announced regular price was one hundred and thirty-five pounds.

The price had gone up while they waited. It was more than two and a half times the original price.

It was not a person who set this price, but an algorithm. And the data the algorithm saw was like this. Hundreds of thousands of people are currently standing in line. These people have been looking at the screen for hours. They will not leave.

The time waited was calculated as evidence of desperation.

It is a structure where the longer a person waits, the more expensively they buy. They build something like this and call it supply and demand.

The British competition authority launched an investigation. The same thing happened at Paul McCartney and Bruce Springsteen concerts. The Royal Opera House raised ticket prices to four hundred and fifteen pounds using this method and received criticism.

The 2026 World Cup organizing committee also introduced this method for tickets and faced backlash from soccer fans.

It is now common for prices to be set differently for each person.

The travel booking site Orbitz looked at the connected computers and showed different results. For people using Macs, more expensive hotels appeared at the top. The calculation was that using an expensive computer means they have a lot of money.

The American education company Princeton Review priced its online courses differently according to zip codes. They sold the same course at a higher price to students in areas where many Asian Americans live.

The data showing that they are a group with high enthusiasm for education was reflected in the price.

The dating app Tinder set its paid subscription fees much higher for older users and sexual minority users, which resulted in legal sanctions and a settlement.

The grocery delivery app Instacart set different prices for each user for the same product at the same supermarket. While supplying the same school supplies to various schools, they also set different prices for each school.

The Washington Post received criticism after introducing a method of setting different subscription fees for each reader.

The American insurance company Allstate created a list of customers who have a tendency not to complain even if rates go up, and charged them more expensive premiums. A case where up to 30 percent higher premiums were charged to residents of areas where people of color and low-income earners live was also mentioned earlier. British car insurance companies were also caught charging higher premiums in areas where many racial minorities live.

Here, we must point out the core of this method.

In the existing market, the price was attached to the product. When you go into a store, there is a price tag, and the person next to you pays the same amount. If it is expensive, you go to another store.

Personalized pricing overturns that structure. The price is attached to the person, not the product.

Then comparison becomes impossible. This is because you cannot know how much the person next to you paid. There is no way to check if you have been ripped off.

For the market to work, it must be possible to compare prices. That condition is disappearing.

Even inside the store, prices move.

The large American retailer Kroger attached electronic price tags and a pricing algorithm. Because there is a small screen instead of paper, the price can be changed at any time.

On a hot day, the price of ice cream goes up. During times when people crowd, the price of drinks goes up.

Senators warned that this technology could raise the prices of essential goods depending on the time of day, weather, and events. Electronic price tags have been confirmed at Whole Foods, Amazon Fresh, Kroger, and Walmart.

Delta Air Lines announced to its shareholders that it is making pricing more profitable with customer data and artificial intelligence.

The fast food chain Wendy's announced that it would introduce time-based pricing at its drive-thrus, but dropped the idea after facing a barrage of public criticism.

The American government is also making a move.

The Federal Trade Commission sent subpoenas to and received data from eight companies that use algorithms to categorize individuals and set targeted prices. They asked what data they get from where, how the algorithms operate, and what impact they have on the prices consumers pay.

In a congressional testimony in April 2026, the leadership of the Federal Trade Commission stated that they are continuing this work and are reviewing whether additional disclosures are needed when prices become personalized.

On March 5, 2026, the House Oversight Committee officially began an investigation. They sent letters to major travel agencies and platform companies requesting materials on revenue management algorithms, the use of consumer data, experimental practices, and internal communications.

In disaster situations, this method becomes even more cruel.

In 2014, when a hostage situation occurred in Sydney, Australia, citizens trying to get out of the city center turned on Uber. Fares jumped to 400 percent of the usual rate.

The same thing happened when Hurricane Sandy passed through the eastern United States in 2012, during the New York City subway shooting in 2022, and immediately after the London terror attacks in 2017.

The algorithm does not know what has happened. It only reads the signal that people are crowding.

A person made the rule to read that signal and raise the price. They did not make a device to turn this rule off during a disaster.

According to a study, Uber's pricing algorithm raised passenger fares while actually reducing the share going to the drivers. It is a structure that takes a little bit from both sides. Neither the passenger nor the driver knows how much the other person paid or received.

Even for housing prices, algorithms match the price.

A rent calculation system created by an American company gathered rent data from millions of households and informed landlords of recommended rents.

Originally, in the rental market, landlords compete with each other. This is because it is better to rent it out, even if it is a little cheap, than to leave it empty.

However, if everyone follows the recommended price of the same algorithm, competition disappears. No one lowers the price.

When people gather and match prices, it is collusion. If an algorithm matches the price, the law has not yet decided what to call it.

Amazon's Project Nessie was also on the same principle. They subtly raise the price of their product and watch to see if competitors follow suit. If they follow, they maintain that price, and if they do not follow, they lower it.

Because machines read each other's signals, the price of the entire market goes up.

The order of search results is also money.

In October 2020, the Korean Fair Trade Commission imposed a 267 billion won fine on Naver. The company had modified its search algorithm to promote its own shopping products and videos to the top while pushing competitor platform products down.

People think of search results as information. They believe rankings reflect popularity or relevance.

That those rankings are commercial products is not written on the screen.

Coupang also faced sanctions for adjusting its algorithm to promote its own brand products and mobilizing thousands of employees to write favorable

reviews and ratings.

Star ratings are a mechanism that makes people believe they represent other consumers' experiences. The company exploited that belief.

Amazon's Buy Box also became a target of regulators. When a customer clicks the add-to-cart button, an algorithm determines which seller they purchase from.

Amazon gave bonus points to sellers using its logistics service and excluded cheaper independent sellers from recommendations. Calculations showed that British consumers alone paid over a billion pounds more.

It sold more expensively by not showing customers cheaper options.

An Indian online marketplace company sued ChatGPT for completely removing its store information from search results. China's Tencent faced criticism for blocking competitor links from opening in its messaging app.

In politics and public opinion, algorithmic manipulation becomes an even greater problem.

X, acquired by Elon Musk, faced analysis from British and American researchers showing that it changed its algorithm to push content and accounts with certain political leanings more prominently into users' timelines.

Meta's Facebook changed its algorithm claiming it would increase meaningful social interaction. What came to light through whistleblowing was this: it gave five times more weight to anger reactions than to likes.

It created a system where angry posts spread more widely, then asked people why they were angry.

Facebook blocked news articles in response to Australia's news licensing bill, but in doing so also blocked emergency relief organization accounts. In India, it faced criticism for removing hashtags related to farmer protests and

hashtags calling for the prime minister's resignation.

By this point, this entire chapter comes together as a single sentence.

We see rankings every day. Search results, recommendation lists, star ratings, prices.

We do not know how those numbers are determined. When we ask, we are told it is a trade secret.

And those who set the numbers know a great deal about us. What devices we use, where we live, how long we wait, whether we complain when prices rise.

It is a transaction where one side knows the other perfectly while the other side knows nothing. These transactions happen dozens of times a day.

Let me recall the incidents in this book from the beginning.

Rejection emails arriving at dawn. Debt collection notices for debts that don't need to be paid. Handcuffs fastened in front of daughters. A recommendation list for a book that doesn't exist. Sixty-three citations where only six were real. An autonomous vehicle that hit a child in front of a school. A robotic arm that recognized a person as a box. Seven thousand images created from an eleven-year-old's photograph. Ticket prices raised to the amount charged for hours of waiting.

There is a common thread running through these incidents.

In every case, the system worked exactly as designed. It was not broken.

It was built to process quickly, so it processed quickly. It was built to cut costs, so it removed procedures for human verification. It was built to increase profit, so it increased profit.

The machine did what it was told to do well.

In the room where what to command was decided, people were sitting. Those who were not in that room paid the price.

Who is sitting in that room, and who will speak on behalf of those who are not there.

This is a question that has not yet been answered.

When the Algorithm Gets It Wrong: AI Harms and Ethical Questions from the AIAAIC Record

Author | Kim Kyung-jin
Publisher | Kim Kyung-jin
Imprint | Kim Kyung-jin Attorney Press
Price : USD 15

© Kim Kyung-jin 2026
All rights reserved. No part of this book may be reproduced without permission.
Note) AI tools assisted with parts of the research and editing of this book.

kimkj008@gmail.com

Support This AI Library Book

If this book was useful, you can support future translations, public editions, and open AI knowledge projects through PayPal.

PayPal

https://paypal.me/kimkjkj

Scan the QR code or open the link above.