Chapter 25

Taking Stock

But: connecting the world meant that we also connected all the bad things and all the bad people, and now every social and political problem is expressed in software. We've had a horrible 'oh shit' moment of realisation, but we haven't remotely worked out what to do about it.
– Benedict Evans

If you campaign for liberty you're likely to find yourself drinking in bad company at the wrong end of the bar.
– Whit Diffie

25.1 Introduction

Our security group at Cambridge runs a blog, `www.lightbluetouchpaper.org`, where we discuss the latest hacks and cracks. Many of the attacks hinge on specific applications, as does much of the cool research. Not all applications are the same, though. If our blog software gets hacked it will just give a botnet one more server, but there are other apps from which money can be stolen, others that people rely on for privacy, others that mediate power, and others that can kill.

I've already discussed many apps from banking through alarms to prepayment meters. In this chapter I'm going to briefly describe four classes of application at the bleeding edge of security research. They are where we find innovative attacks, novel protection problems, and thorny policy issues. They are: autonomous and remotely-piloted vehicles; machine learning, from adversarial learning to more general issues of AI in society; privacy technologies; and finally, electronic elections. What these have in common is that while previously, security engineering was about managing complexity in technology with all its exploitable side-effects, we are now bumping up against complexity in human society. Autonomous cars are hard because of the people driving other cars on the same road. AI is hard because our cool new pattern-matching tools, such as deep neural networks, can pick out not just real patterns in human behaviour –

sometimes unexpected ones – but false ones too. Privacy is hard because of the richness of human interaction in society. And elections are hard not just because of the technical difficulty of counting votes in a way that preserves both privacy and auditability, but because of the huge variety of dirty tricks used by political players, both upstream and downstream of the voting process itself. All of these problems explore, in various ways, the boundary between what humans can do and what machines can do.

25.2 Autonomous and remotely-piloted vehicles

The aviation pioneer Lawrence Sperry invented the first autopilot in 1912 and demonstrated it in 1914, flying past the judges in a 'safer aircraft' competition in Paris with his hands up. In the process he and his father Elmer invented the artificial horizon. A fixed-wing aircraft left to its own devices will eventually go into a spiral dive and crash; the pilot can keep it level with reference to the horizon, but when flying in cloud that external reference is missing. A gyroscope can provide the missing reference, and it can also drive the ailerons and elevators via servos.

In 1975, I got my first proper job re-engineering a fast-jet inertial navigation set to work on the midget submarines used in the oil industry. Engineers in the same building were working on early head-up displays and satellite navigation equipment. Each of these pieces of equipment weighed about 20kg and cost £250,000 – about $3M in today's money. All three together left little change out of $10M and weighed as much as a person.

Now, in 2020, you have all three in your phone. Rather than three spinning mechanical gyros in a precision-engineered cage, your phone has a chip with MEMS accelerometers and gyros. It also has a GPS chip for satellite navigation and a Google or Apple Maps app to show you how to walk, cycle or drive to your destination. Over forty years, the cost has fallen by six orders of magnitude and the mass by four. This has driven rapid evolution of assistive technology on sea, air and land. Pioneering single-handed yachtsmen developed self-steering gear to cross oceans from the 1920s, to give them time to sleep, cook and repair sails; amateurs now have smarter autopilots for coastal cruising. Autonomous probes swim beneath the Antarctic ice to measure how quickly it's melting. The world's navies develop underwater mines, autonomous submersibles to find them, and much else.

25.2.1 Drones

In the air, early weapons such as the German V1 and V2 used twin-gyro autopilots, while the Cold War gave us the Tomahawk cruise missiles used to great effect in both Gulf Wars. In service since the early 1980s, these sneak under the enemy radar by flying close to the ground, and use terrain contour matching to update their inertial navigation. They were followed closely by a variety of unmanned aerial vehicles (UAVs), which saw their first large-scale use in the war between Israel and Syria in 1982; the Israeli Air Force used them for reconnaissance and as decoys, wiping out the Syrian air force with minimal losses. The

best-known of the next generation of UAVs was the Predator. Initially designed as a reconnaissance vehicle, it could linger over a target area at medium altitude for many hours, and was adapted to carry Hellfire missiles to strike targets on the ground. In service from 1995–2018, it saw service in Iraq, Afghanistan, Libya and elsewhere. It was replaced by the larger, faster Reaper, which became a mainstay of the war in Syria against Islamic State. The world's armed forces now have a large range of UAVs, right down to small drones that soldiers carry in their rucksacks and use to see what's round the next corner.

Through the 20th century, enthusiasts built small radio-controlled model aircraft, but the FAA only issued its first commercial drone permit in 2006. In 2010, Parrot unveiled its AR Drone, a quadcopter that could be controlled by wifi from a smartphone, and in 2013 Amazon announced it was considering drones for delivery. Interest took off rapidly; within a couple of years our students were building drones and soon you could buy low-cost models in hobby shops. The main application in 2020 is aerial photography. There have been both insurgent and criminal uses, though, with drones used to deliver both drugs and mobile phones to prisoners, while insurgents have fitted drones with improvised explosive devices for use as weapons.

25.2.2 Self-driving cars

Most of the recent surge in interest though has been in self-driving cars and trucks. In 2004, faced with mounting combat losses to improvised explosive devices in Afghanistan and Iraq, DARPA decided to push the development of self-driving vehicles, and announced a competition with a million-dollar prize for whoever built one that could cross 149 miles of the Mojave desert the fastest. The prize went unclaimed as no vehicle finished the course, but the following year a team from Stanford led by the roboticist Sebastian Thrun collected the prize, now two million. His robot, Stanley, used machine learning and probabilistic reasoning to cope with terrain perception, collision avoidance, and stable vehicle control on slippery and rugged terrain [1887]. This built on robotics research going back to the 1980s, much of which DARPA had also funded. Their next challenge in 2007 moved from the desert to a simulated urban environment; competitors had to detect and avoid other vehicles, and obey the rules of the road. This bootstrapped a research community and the technology started to improve quickly.

Previously, carmakers had been steadily adding assistive technology, starting with antilock braking systems (ABS) in the last century and progressing through adaptive cruise control (ACC), which I described in section 23.4.1, automatic emergency braking (AEB) and lane keeping assist (LKA). The industry vision was that these would eventually come together into a full autopilot. Inspired by the DARPA challenges, Google hired Sebastian Thrun to lead Project Chauffeur in 2009 with a goal of building a fully self-driving car. This was announced in 2010, stimulating a market race involving both the tech and auto industries. Tesla was the first to field a product in 2014, when its 'Autopilot' software was launched as an over-the-air upgrade that could take control on the freeway or in stop-start traffic. There was already a hype cycle underway for machine learning, which I'll discuss in the next section, and self-driving cars hitched a

ride. Tesla's Elon Musk was predicting full autonomy by 2018, and Google's Sergey Brin by 2017, before the Google car project was spun off as Waymo in 2016. People talked excitedly about low-cost robotaxis causing personal car ownership to be replaced by mobility-as-a-service; the arrival of Uber added a further competitor, and the hype scared even auto industry execs who should have known better into predicting that by the mid-2020s people wouldn't own their own cars any more. The hype cycle passed, as it always does. As I write in 2020, Waymo is operating a limited self-driving car service in a 50-square-mile area of Phoenix [871]. The service isn't available when it's raining, or in dust storms, and is monitored in real-time by humans at a control centre. It had been announced several times, but problems kept on forcing the company to put safety drivers back in the cars. So what's going on?

A large part of the answer is that other road users are unpredictable. Automation can deal with some of the resulting hazards: if the car in front brakes suddenly, a robot can react faster. Adaptive cruise control cuts driver fatigue and even cuts congestion once enough vehicles use it, as it damps the propagation of shock waves through traffic. But even here there are limits. When engineers extended the technology to automatic emergency braking, the inability to infer the intentions of other drivers became a limiting factor. Suppose for example you're driving on an open country road when the car in front indicates a turn and starts to slow down. You maintain speed as you expect it'll have left the road by the time you get there, and if not you'll just overtake. But the AEB might not understand this, so as you get too close to the turning car it activates, throwing you forward on your seat belt. Consumer tests of AEB systems in 2020 still show quite some variability, both in the false alarm rate and in the ability to stop the car in time when a pedestrian dummy is pulled across the road. Some systems restrict activation to city rather than highway speeds, and in 2020 all tend to be options available on more expensive models. AEB should be in all new cars in about 2022. Since 2016 insurers have been happy that it reduces the overall risk; I'll discuss safety assurance in section 28.4.1.

But each new assistive technology takes years to optimise and debug, and it's not straightforward to combine a dozen of them into an autopilot. The paper that Sebastian Thrun and his team wrote to describe Stanley gives a useful insight into the overall technology [1887]. There are several dozen programs interacting loosely, reflecting our understanding of how humans do such tasks; your subconscious looks at all sorts of things and brings hazards to your attention. Simultaneous processes in Stanley handled path planning, steering control and obstacle avoidance; this used laser rangefinders up to 22m, a colour camera beyond that, and a radar beyond that (which was not used in the race, as Stanley was given over 2000 waypoints for a predetermined course). Each of these systems had to solve many subproblems; the vision system, for example, had to adapt to changing light conditions and road colour. Stanley then had to be optimised via extensive testing, where the objective function was to maximise the mean distance between catastrophic failure (defined as the human safety driver taking over).

Combining the subsystems means compromises, and while the main vendors hold their design details secret, we're starting to learn about the optimisations and what goes wrong with them from accidents. For example, when a self-

driving Uber killed Elaine Herzberg in Arizona in March 2018, it emerged at the NTSB inquiry that Elaine had been pushing a bicycle and the vision system flapped between identifying her as a pedestrian and as something else, but ultimately she was not recognised as a pedestrian because she was not on a crosswalk. AEB might have stopped the car but it had been turned off "to reduce the potential for erratic vehicle behavior" – in other words, because the false alarm rate was annoying [457]. Ultimately, Uber relied on the safety driver – who was unfortunately watching TV at the time[1].

Now we've known for decades that relying on humans to take over in an emergency takes time: a human has to react to an alarm, analyse the alarm display on the console, scan the environment, acquire situational awareness, get into the optical flow, and take effective control. Even in commercial aviation, it takes a flight crew about eight seconds to regain control properly after an autopilot failure. You cannot expect a safety driver in a car to do much better.

25.2.3 The levels and limits of automation

For such reasons, the Society of Automotive Engineers sets out five levels of automation:

1. Driver assistance – the software controls either steering or speed, and the human driver does the rest of the work;
2. Partial automation – the software controls both steering and speed in some modes but the human driver is responsible for monitoring the environment and assuming control at zero notice if the software gets confused;
3. Conditional automation – the software monitors the environment, and controls both steering and speed, but assumes the human can take over if it gets confused;
4. High automation – the software monitors the environment and drives the car, in some driving conditions, without assuming that a human can intervene. If it gets confused it stops at the side of the road;
5. Full automation – the software can do everything a human can.

So far, vehicles available on the mass market only have *advanced driver assistance systems* (ADAS), namely levels one and two, and insurers consider words like 'autonomous' and 'autopilot' to be dangerous as they cause customers to assume that a vehicle is operating at Level 4, which can lead to accidents. The Arizona crash can be seen as a car operating at Level 2 while the safety driver operated at Level 3. Level 4 often assumes a backup driver sitting in a control centre, overseeing several dozen 'autonomous' cars, but they won't have the bandwidth to understand a hazard as quickly as a safety driver on the spot. They don't feel the road noise and accelerations, they can't use their peripheral vision, and above all, they are not immersed in the optical flow field

[1] In fact, the very first fatal crash involving a Tesla on autopilot claimed the life of a driver who appeared to be watching a movie on his laptop when his car ran under a truck [1394].

that is critical to driving a car (or landing an aircraft) safely, as we discussed in section 3.2.1.

To what extent is Level 5 feasible at all, unless we invent artificial general intelligence? John Naughton remarked that a downtown delivery driver's job is pretty safe, as the work demands all sorts of judgment calls such as whether you can double-park or even block a narrow street for half a minute while you dash up to a doorway and drop a parcel, as the cars behind honk at you [1417]. Another hard case is the cluttered suburban street with cars parked either side, where you are forever negotiating who goes first with oncoming vehicles, using a wave, a nod or even just eye contact. Even the current Level 2 systems tend to have difficulty when turning across traffic because of their inability to do this tacit negotiation. They end up having to be much more cautious than a human driver and wait for a bigger gap, which annoys human drivers behind them. And if you've ever tried to ease a car through the hordes of students on bicycles in a college town like Cambridge, or any urban traffic in India, you know that dealing with human traffic complexity is hard in many other situations. Can your self-driving car even detect hand signals from police officers to stop, let alone cope with eight students carrying a bed, or with an Indian temple procession?

As of 2020, the Level 2 systems have lots of shortcomings. Tesla can't always detect stationary vehicles reliably; it uses vision, sonar and radar but no lidar. (One Tesla driver in North Carolina has been charged after running into the back of a stationary police car [1118].) The Range Rover can't always detect the boundary between a paved road and grass, but perhaps that wasn't a priority for a 4 x 4. Many cars have issues with little roundabouts, not to mention potholes and other rough surfaces; the first time I got a ride in one, my teeth were rattled as we went over speed bumps at almost 30mph. Roadworks play havoc with automatic lane-keeping systems, as old white lines that have been painted over can be shiny black and very prominent in some light conditions, leading cars to oscillate back and forth between old and new markings [632]. There's a huge amount of research on such technical topics, from better algorithms for multi-sensor data fusion though driving algorithms that can provide an explanation for their decisions, to getting cars to learn routes as they travel them, just like humans do. Tesla even has a 'shadow mode' for its autopilot; when it's not in use, it still tries to predict what the driver will do next, and records its mispredictions for later analysis. This has enabled Tesla to collect billions of miles of training data across a vast range of road and weather conditions.

I'll discuss safety assurance in section 28.4.1 but the state of play in 2020 is that while Tesla and NHTSA claimed that there are fewer crashes after a Tesla customer activates Autosteer, an independent lab claimed there were more. Now as I discussed in section 14.3.1, falling asleep at the wheel is a major cause of accidents, accounting for 20% of the UK total. These tend to be at the serious end of the spectrum; they account for about 30% of fatal accidents and half of fatal accidents on freeways. (That's why we have laws to limit commercial drivers' hours.) So we ought to be able to save lives with a system that keeps your car in lane on the freeway, brakes to avoid collisions, and brings it to a stop at the side of the road if you don't respond to chimes. Why is this not happening?

I suspect we'll need to disentangle at least three different factors: the risk

thermostat, the system's affordances, and the expectations created by marketing. First, the risk thermostat is the mechanism whereby people adapt to a perceived reduction in risk by adopting more risky behaviour; we noted in section 3.2.5.7 that mandatory seat-belt laws caused people to drive faster, so that the overall effect was merely to move casualties from vehicle occupants to pedestrians and cyclists, rather than to reduce their number overall. Second, affordances condition how we interact with technology, as we discussed in section 3.2.1, and if a driver assistance system makes driving easier, and apparently safer, people will relax and assume it is safer – disposing some of them to take more risks. Third, the industry's marketing minimises the risks in subtle ways. For Tesla to call its autosteer feature an autopilot misled drivers to think they could watch TV or have a nap. That is not the case with an autopilot on an airplane, but most non-pilots don't understand that.

25.2.4 How to hack a self-driving car

The electronic security of road vehicles started out in the last century with the truck tachographs and speed limiters we discussed in section 14.3 and the remote key entry systems we discussed in section 4.3.1. It has become a specialist discipline since about 2005, when the carmakers and tier-1 component vendors started to hire experts. By 2008, people were working on tamper resistance for engine control units: the industry had started using software to control engine power output, so whether your car had 120 horsepower or 150 was down to a software switch which people naturally tried to hack. The makers tried to stop them. They claimed they were concerned about the environmental impact of improperly tuned cars, but if you believe that, I have a bridge I'd like to sell you.

In 2010, Karl Koscher and colleagues got the attention of academics by showing how to hack a late-model Ford. Cars' internal data communications use a CAN bus which does not have strong authentication, so an attacker who gets control of (say) the radio can escalate this access to operate the door locks and the brakes [1085]. In 2015, Charlie Miller and Chris Valasek got the attention of the press when they hacked a Jeep Cherokee containing a volunteer journalist, over its mobile phone link, slowed the vehicle down and drove it off the road [1316]. This compelled Chrysler to recall 1.4m vehicles for a software patch, costing the company over $1bn. This finally got the industry's attention.

There's now a diverse community of people who hack cars and other vehicles. There are hobbyists who want to tune their cars; there are garages who also want to use third-party components and services; and there are farmers who want to repair their tractors despite John Deere's service monopoly, as I mentioned in section 24.6. There are open-source software activists and safety advocates who believe we're all safer if everything is documented [1792]. And there are the black hats too: intelligence agencies that want to spy on vehicle occupants and thieves who just want to steal cars.

Car theft is currently the main threat model, and we discussed the methods used to defeat remote key entry and alarm systems in section 4.3.1. State actors and others can take over the mobile phones embedded in cars, using the techniques discussed in section 2.2.1. The phones, navigation and infotainment

systems are often poorly designed anyway – when you rent a car, or buy one secondhand, you often see a previous user's personal information, and we described in section 22.3.3 how an app that enables you to track and unlock a rental car let you continue to do this once the car had been rented to somebody else.

So what else might go wrong, especially as cars become more autonomous? A reasonable worst-case scenario might see a state actor, or perhaps an environmental activist group, trying to scare the public by causing thousands of simultaneous road traffic accidents. A remote exploit such as that on the Chrysler Jeep might already do this. The CAN bus which most modern cars use for internal data communications trusts all its nodes. If one of them is subverted it might be reprogrammed to transmit continuously; such a 'blethering idiot', as it's called, makes the whole bus unusable. If this is the powertrain bus, the car becomes almost undriveable; the driver will still have some steering control but without power assistance to either steering or brakes. If the car is travelling at speed, there's a serious accident risk. The possibility that a malicious actor could hack millions of cars causing tens of thousands of road traffic accidents simultaneously is unacceptable, and such vulnerabilities therefore have to be patched. But patching is expensive. The average car might contain 50–100 electronic control units from 20 different vendors, and the integration testing needed to get them to all work together smoothly is expensive. I'll discuss this in more detail in section 27.5.4.

Attacks are not limited to the cars themselves. In 2017, Elon Musk told an audience, "In principle, if someone was able to say hack all the autonomous Teslas, they could say – I mean just as a prank – they could say 'send them all to Rhode Island' – across the United States ... and that would be the end of Tesla and there would be a lot of angry people in Rhode Island.". His audience laughed, and three years later it emerged that he'd not been entirely joking. A few months previously, a hacker had gained control of the Tesla 'mothership' server which controls its entire fleet; luckily he was a white hat and reported the hack to Tesla [1119]. At the other end of the scale, the performance artist Simon Weckert pulled a handcart containing 99 Android phones around Berlin in February 2020, causing Google Maps to register a traffic jam wherever he went [1997]. As advanced driver assistance systems rely ever more extensively on cloud facilities, the scope for such indirect attacks will increase.

And external attacks need not involve computers. If car systems start to slow down automatically for pedestrians and cyclists, some of them may exploit this. In India and some parts of southern Europe, pedestrians walk through congested traffic, flagging cars to stop, and they do; it will be interesting to see if this behaviour appears in London and New York as well.

Companies will exploit assistance systems if they can. Now that the initial dream of self-driving trucks seems some way off, and even the intermediate dream of multiple trucks driving in convoy between distribution hubs with a single driver seems ambitious, may we expect lobbying to relax the legal limits on drivers' hours? Trucking firms may argue that once the truck's on autopilot on the freeway, the driver only has to do real work on arrival and departure, so he should work ten hours a shift rather than eight. But if the net effect of the technology is to make truck drivers work more time for the same money, it will

be resented and perhaps sabotaged.

Should Level 5 automation ever happen, even in restricted environments – so that we finally see the robotaxis Google hoped to invent – then we'll have to think about social hacking as a facet of safety. If your 12-year-old daughter calls a cab to get a ride home from school, then at present we have safeguards in the form of laws requiring taxi drivers to have background checks for criminal records. Uber tried to avoid these laws, claiming it wasn't a taxi company but a 'platform'; in London, the mayor had to ban them and fight them in court for years to get them to comply. So how will safeguarding work with robotaxis?

There will also be liability games. At present, car companies try to blame drivers for crashes, so each crash becomes a question of which driver was negligent. If the computer was driving the car, though, that's product liability, and the manufacturer has to pay. There have been some interesting tussles around the safety figures for assisted driving, and specifically whether the carmakers undercount crashes with autopilot activated, which we'll discuss in section 28.4.1.

So much is entirely predictable. But what about new attacks on the AI components of the systems themselves? For example, can you confuse a car by projecting a deceptive image on a bridge, or on the road, and cause it to crash? That's quite possible, and I've already seen a crash caused by visual confusion. On the road home from my lab, there was a house at a right-hand bend whose owner often parked his car facing oncoming traffic. At night, in a left-hand driving country like Britain, your driving reflex is to steer to the left of the facing car, but then you'd notice you were heading for his garden wall, and swerve right to pass to the right of his car instead. Eventually a large truck didn't swerve in time, and ended up in the wall.

So could clever software fool a machine vision system in new ways, or ways that might be easier for an attacker to scale? That brings us to the next topic, artificial intelligence, or to be more precise, machine learning.

25.3 AI / ML

The phrase *artificial intelligence* has meant different things at different times. For pioneers like Alan Turing, it ranged from the Turing test to attempts to teach a computer to play chess. By the 1960s it meant text processing, from Eliza to early machine translation, and programming in Lisp. In the 1980s there was a surge of research spurred by Japan's announcement of a huge research programme into 'Fifth generation computing', with which Western nations scrambled to keep up; much of that effort went into rule-based systems, and Prolog joined Lisp as one of the languages on the computer science curriculum.

From the 1990s, the emphasis changed from handcrafted systems with lots of rules to systems that learn from examples, now called *machine learning* (ML). Early mechanisms included logistic regressions, support vector machines (SVMs) and Bayesian classifiers; progress was driven by applications such as natural language processing (NLP) and search. While the NLP community developed custom methods, the typical approach to designing a payment fraud detector or spam filter was to collect large amounts of training data, write custom code

to extract a number of signals, and just see empirically which type of classifier worked best on them. Search became intensely adversarial during the 2000s as search engine optimisation firms used all sorts of tricks to manipulate the signals on which search engines rely, and the engines fought back in turn, penalising or banning sites that use underhand tricks such as hidden text. Bing was an early user of ML, but Google avoided it for years; the engineer who ran search from 2000 until he retired in 2016, Amit Singhal, felt it was too hard to find out, for a given set of results, exactly which of the many inputs was most responsible for which result. This made it hard to debug machine-learning based algorithms for search ranking. If you detected a botnet clicking on restaurants in Istanbul and wanted to tweak the algorithm to exclude them, it was easier to change a few 'if' statements than retrain a classifier [1300].

A sea change started in 2011 when Dan Cireşan, Ueli Meier, Jonathan Masci and Jürgen Schmidhuber trained a deep convolutional neural network to do as well as humans on recognising handwritten digits and Chinese characters, and better than humans on traffic signs [435]. The following year, Alex Krizhevsky, Ilya Sutskever and Geoff Hinton used a similar *deep neural network* (DNN) to get record-breaking results at classifying 1.2 million images [1098]. The race was on, other researchers piled in, and 'deep learning' started to get serious traction at a variety of tasks. The most spectacular result came in 2016 when David Silver and colleagues at Google Deepmind produced AlphaGo, which defeated the world Go champion Lee Sedol [1737]. This got the attention of the world. Before then, few research students wanted to study machine learning; since then few want to study anything else. Undergraduates even pay attention in classes on probability and statistics, which were previously seen as a chore.

25.3.1 ML and security

The interaction between machine learning and security goes back to the mid-1990s. Malware writers started using tricks such as polymorphism to evade the classifiers in anti-virus software, as I described in section 21.3.5; banks and credit card companies started using machine learning to detect payment fraud, as I described in section 12.5.4; and phone companies also used it for first-generation mobiles, as I noted in section 22.2. The arrival of spam as the Internet opened up to the public in the mid-1990s created a market for spam filters. Hand-crafted rules didn't scale well enough for large mail service providers, especially once botnets appeared and spam became the majority of email, so spam filtering became a big application.

Alice Hutchings, Sergo Pastrana and Richard Clayton surveyed the use of machine-learning in such systems, and the tricks the bad guys have worked out to dupe them [939]. As spam filtering takes user feedback as its ground truth, spammers learned to send spam to accounts they control at the big webmail firms, and mark it 'not spam'; other statistical analysis mechanisms are now used to detect this. Poisoning a classifier's training data is a quite general attack. Another is to look for weak points in a value chain: airline ticket fraudsters buy an innocuous ticket, pass the fraud checks, and then change it just before departure to a ticket to a high-risk destination. And there are vigorous discussions of such techniques on the underground forums where the

bad actors trade not just services but boasts and tips. Battista Biggio and Fabio Rolli give more technical background: in 2004, spammers found they could confuse the early linear classifiers in spam filters by varying some of the words, and an arms race took off from there [241].

It turns out that these attack ideas generalise to other systems, and there are other attacks too.

25.3.2 Attacks on ML systems

There are at least four types of attack on a machine-learning system.

First, you can poison the training data. If the model continues to train itself in use, then it might be simple to lead it astray. Tay was a chatbot released by Microsoft in March 2016 on Twitter; trolls immediately started teaching it to use racist and offensive language, and it was shut down after only 16 hours.

Second, you can attack the model's integrity in its inference phase, for example by causing it to give the wrong answer. In 2013, Christian Szegedy and colleagues found that the deep neural networks which had been found to classify images so well in 2012 were vulnerable to *adversarial samples* – images perturbed very slightly would be wildly misclassified [1857]. The idea is to choose a perturbation that maximises the model's prediction error. It turns out that neural networks have plenty of such blind spots, which are related to the training data in non-obvious ways. The decision space is high-dimensional, which makes blind spots mathematically inevitable [1706]; and with neural networks the decision boundaries are convoluted, making them non-obvious. Researchers quickly came up with real-world adversarial examples, ranging from small stickers that would cause a car vision system to misread a 30mph speed sign as 60mph, to coloured spectacles that would cause a man wearing them to be misrecognised as a woman, or not recognised at all [1720]. In the world of malware detection, people found that non-linear classifiers such as SVM and deep neural networks were not actually harder to evade than linear classifiers provided you did it right [241].

Third, Florian Tramèr and colleagues showed that you can attack the model's confidentiality in the inference phase, by getting it to classify a number of probe inputs and building a successively better approximation. The result is often a good working imitation of the target model. As in the manufacture of real goods, a knock-off is often cheaper; big models can cost a lot to train from scratch. This approximation attack works not just with neural networks but also with other classifiers such as logistic regression and decision trees [1901].

What's more, many attacks turn out to be transferable, so an attacker doesn't need full access to the model (a so-called *white-box attack*) [1900]. Many attacks can be developed on one model and then launched against another that's been trained on the same data, or even just similar data (a *black-box attack*). The blind spots are a function of the training data, so in order to make attacks less transferable you have to make an effort. For example, Ilia Shumailov, Yiren Zhao, Robert Mullins and I have experimented with inserting keys in neural networks so that the blind spots appear in different places, and models with different keys are vulnerable to different adversarial samples [1733]. Kerckhoffs'

principle applies in machine learning, as almost everywhere else in security.

A variant on the confidentiality attack is to extract sensitive training data. Large neural networks contain a lot of state, and the simplest way to deal with outliers is often just to memorise them. So if some business claims that a classifier trained on a million medical records is not personal data because it's "statistical machine learning", take care. Ways of combining machine learning with differential privacy, which we discussed in section 11.3, are a subject of active research [1493].

Finally, you can deny service, and one way is to choose samples that will cause the classifier to take as long as possible. Ilia Shumailov and colleagues found that one can often deny service by posing a conundrum to a classifier. Given a straight-through pipeline, as in a typical image-processing task, a confusing image can take 20% more time, but in more complex tasks such as natural language processing you can invoke exception handling and slow things down hundreds of times [1730].

More complex attacks straddle these categories. For example, there's an arms race between online advertisers and the suppliers of ad-blocking software, and as the advertisers adopt ever more complicated ways of rendering web pages to confuse the blockers, the blockers are starting to use image processing techniques on the rendered page to spot ads. However this leaves them open to advertisers using adversarial samples either to escape the filter, or to cause it to wrongly block another part of the page [1899].

So how can one use machine learning safely in the real world? That's something we're still learning, but there are some things we can say. First, one has to take a systems security approach and look at the problem end-to-end. Just as we sanitise inputs to web services, do penetration testing, and have mechanisms for responsible disclosure and update, we need to do the same for ML systems [659].

Second, we need to draw on the experience of the last twenty years' work on topics like card fraud, spam and intrusion detection. As we mentioned in section 21.4.2.2, ML systems have been largely ineffective at real-world network intrusion detection; Robin Sommer and Vern Paxson were the first to give a good explanation why. They discuss the lack of training data, the distance between theory and practice, the difficulties in evaluation, the high cost of errors and above all the inability to deal with novel attacks [1802]. The problem of keeping capable opponents out of complex corporate networks just isn't one that artificial intelligence has ever been good at.

There may occasionally be a change in emphasis, though. If we want to lower the probability of a new adversarial attack causing real damage, there are various things we can do, depending on the context. One is simply to detune the classifier; this is the approach in at least one machine-vision system used in cars. By making it less sensitive, you make it less easy to spoof, and then you complement it with other sensors such as radar and ultrasonics so that the vision system on its own is less critical. An alternative approach is to head in the other direction, by making the ML component of your system sufficiently fragile that an attack can be detected by other components – whereupon you switch to a defensive mode of operation, such as a low-sensitivity limp-home mode or

stopping and waiting for a human to drive. In other words, you set out to build in situational awareness. This is how we behave in real life; as I discussed in section 3.2.5.1, the ancestral evolutionary environment taught us to take extra care when we sense triggers such as adversarial intent and violations of tribal taboos. So we've experimented with using neural networks trained so that a number of outputs and activations are considered to be taboo and avoided; if any of these taboos is broken, an attack can be suspected [1733].

The fundamental problem is that once we start letting machine learning blur the boundary between code and data, and systems become data-driven, people are going to game them. This brings us to the thorny problem of the interaction of machine learning and society.

25.3.3 ML and society

The surge of interest in machine learning since 2016, and its representation as 'artificial intelligence' in the popular press, has led to a lot of speculation about ethics. For example, the philosopher Dan Dennett objects on moral grounds to the existence of persons that are immortal and intelligent but not conscious. But companies already meet that definition! The history of corporate wrongdoing shows that corporations can behave very badly indeed (we discussed some examples in section 12.2.6). The most powerful ML systems belong to corporations such as Google, Amazon, Microsoft and IBM, all of which have had tussles with authority. The interplay between ML, big data and monopoly adds to the thicket of issues that governments need to navigate as they ponder how to regulate tech. One aspect is that the tech majors' ML offerings are now becoming platforms on their own, and used by lots of startups solving specific real-world problems [658].

One cross-cutting issue is prejudice. Aylin Caliskan, a Turkish research student at Princeton, noticed that machine translations from Turkish to English came out with gender bias; although Turkish has no grammatical gender, the English translations of Turkish sentences would assign doctors as 'he' and nurses as 'she'. On further investigation, she and her supervisors Joanna Bryson and Arvind Narayanan found that essentially all machine translation systems in use were not merely sexist, but racist and homophobic too [369]. In fact a large number of natural-language systems based on machine learning inhale the prejudices of their training data. If the big platforms' ML engines can suffuse prejudice through the systems on which hundreds of downstream firms rely, there is definitely a public-policy issue.

A related policy problem is *redlining*. When insurance companies used postcode-level claim statistics to decide the level of premiums, it was found that many minority areas suffered high premiums or were excluded from cover, breaking anti-discrimination laws. I wrote in the second edition of this book in 2008: "If you build an intrusion detection system based on data mining techniques, you are at serious risk of discriminating. If you use neural network techniques, you'll have no way of explaining to a court what the rules underlying your decisions are, so defending yourself could be hard. Opaque rules can also contravene European data protection law, which entitles citizens to know the algorithms used to process their personal data."

A second cross-cutting issue is snake oil, and the AI/ML gold rush has led to thousands of startups, many of them stronger on marketing than on product. Manish Raghavan and colleagues surveyed 'AI' systems used in employment screening and hiring, finding dozens of firms that claim their systems match new hires to the company's requirements. Most claim they don't discriminate, yet as few employers retain comprehensive and accessible data on employee performance, it's entirely unclear how such systems can even be trained, let alone how a firm that used such a system might defend a lawsuit for discrimination [1571]. Applicants quickly learn to game the system, such as by slipping the word 'Oxford' or 'Cambridge' into their CV in white text. A prudent employer would demand more transparent mechanisms, and devise independent metrics to validate their outcomes. Even that is nontrivial, as machine learning can discover correlations that we do not understand.

Arvind Narayanan has an interesting analysis of snake oil in AI [1382]. 'AI' and even 'ML' are generic terms for a whole grab-bag of technologies. Some of them have made real progress, like DNNs for face recognition, and indeed AlphaGo. So companies exploit this hype, slapping the 'AI' label on whatever they're selling, even if its mechanisms use statistical techniques from a century ago. Digging deeper, Arvind argues that machine-learning systems can be sorted into three categories:

1. ML has made real progress on tasks of *perception*, such as face recognition (see section 17.3), the recognition of songs by products like Shazam (see section 24.4.3), medical diagnosis from scans, and speech-to-text – at all of which it has acquired the competence of skilled humans;

2. ML has made some progress on tasks of *judgment*, such as content recommendation and the recognition of spam and hate speech. These have many hard edge cases about which even skilled humans disagree. The systems that perform them often rely on substantial human input – from a billion email users clicking the 'report spam' button to the tens of thousands of content moderators employed by the big tech companies;

3. ML has made no progress on tasks of *social prediction*, such as predicting employee performance, school outcomes and future criminal behaviour. A very extensive study by Matthew Sagalnik and over 400 collaborators has concluded that insofar as life outcomes can be predicted at all, this can be done as well using simple linear regressions based on a handful of variables [1638].

This is a falsifiable claim, so we'll see how accurate it is over time, and if there's a fourth edition of this book in 2030 we'll have a lot more data then. A major theme of research meanwhile will be to look for better ways for people and machines to work together. Intuitively, we want people to do the jobs involving judgment and machines to do the boring stuff; but making that actually work can be harder than it looks. Often people end up being the machine's servants, and according to one VC firm, 40% of 'AI' startups don't actually use ML in any material way; they're merely riding the wave of hype and employ people behind the scenes [1960]. One way or another, there will be lots of bumps in the road, and lots of debates about ethics and politics.

Perhaps the best way to approach the ethics is this. Many of the problems now being discussed in the context of AI ethics arose years ago for research done using traditional statistical methods on databases of personal information. (Indeed, linear regressions have been used continuously for about a century; they've just been rebranded as machine learning.) So our first port of call should be existing law and policy. When we discussed ethics in the context of records-based health and social-policy research in section 10.4.5.1, we observed that many of the issues arose because IT companies and their customers ignored the wisdom that doctors, teachers and others had accumulated over years of dealing with paper-based records. The same mistakes are now being repeated, and excused as before with sales hype around 'innovation' and 'disruption'.

In the case of predicting which children are likely to turn to crime, it's been known for years that such indicators can be deeply stigmatising. In section 10.4.6 we noted that if you tell teachers which kids have had contact with social services, then the teachers will have lower expectations of them. Both child welfare and privacy law argue against sharing such indicators. How much more harmful might it be if clueless administrators buy software that claims to be making predictions using the inscrutable magic that enabled AlphaGo to beat Lee Sedol? As for 'predictive policing', studies suggest that it might just be another way to get the computer to justify a policy of 'round up the usual suspects' [677]. (In section 14.4 we discussed how curfew tags also have this effect.) Similar issues arise with the use of ML techniques to advise judges in bail hearings about whether a suspect poses a flight risk or reoffending risk, and also in sentencing hearings about whether a suspect is dangerous. Such technologies are likely to propagate existing social biases and power structures, and provide lawmakers with an excuse to continue ineffective but populist policies, rather than nudging them to tackle the underlying problems.

ML is nonetheless likely to upset some of the equilibria that have emerged over the years on issues like surveillance, privacy and censorship, as it makes even more powerful tools available to already powerful actors, as well as creating new excuses to revive old abuses. Many countries already restrict the use of CCTV cameras; now that face-recognition systems enable pedestrians to be recognised, do we need to restrict them more? As we saw in section 17.3, a number of cities (including San Francisco) have decided the answer is 'yes'. In section 11.2.5 we discussed how location and social data can now make it very hard to be anonymous, and how people's Facebook data could be mined for political ad targeting. ML techniques make it easier to do traffic analysis, by spotting patterns of communication [1719]; in fact, police and intelligence agencies depend ever more on traffic and social-network analysis of the sort discussed in sections 21.7, 23.3.1 and 26.2.2.

In short, the charge sheet against machine learning is that it is one of the technologies helping entrench the power of the tech majors while pushing the balance between privacy and surveillance towards surveillance and facilitating authoritarian government in other ways. It may be telling that Google and Microsoft are funding big research programs to develop AI for social good.

So what can we do as a practical matter to get some privacy in this electronic village in which we now live?

25.4 PETS and operational security

Even if you don't blurt out all your thoughts on Facebook, social structure – who hangs out with whom – says an awful lot, and has become much more visible. In section 11.2.5 we discussed research which suggested that as few as four Facebook likes enable a careful observer to work out whether you're straight or gay most of the time, and how this observation led among other things to the Cambridge Analytica scandal, where voters' preferences were documented covertly and in detail.

Even if you don't use Facebook at all, the traffic data on who contacted whom gives a lot away to those who have access to it, as we discussed in section 11.4.1. This can cause problems for people who are in conflict with authority, such as whistleblowers. Anonymity can sometimes be a useful tool here. The abuse of academic authority is countered by anonymous student feedback on professors and anonymous refereeing of conference paper submissions. If your employer pays your health insurance, you might want to buy an HIV test kit for cash and get the results anonymously online, as the mere fact that you took a test says something, even if the result is negative. Privacy can also be a necessary precursor of free speech. People trying to innovate in politics or religion may need to develop their doctrine and build their numbers before going public. And then there are opposition politicians digging a bear trap for the government of the day, whose concerns are more tactical.

The importance of such activities to an open society is such that we consider privacy and freedom of speech to be interlinked human rights. We also enact laws to protect whistleblowers. But how can this work out in practice?

In pre-technological societies, two people could walk a short distance away from everyone else and have a conversation that left no hard evidence of what was said. If Alice claimed that Bob had criticised the king, then Bob could always claim the converse – that it was Alice who'd proposed a demonstration to increase the powers of parliament and he who'd refused out of loyalty.

In other words, many communications were *deniable*. Plausible deniability remains an important feature of some communications today, from everyday life up to the highest reaches of intelligence and diplomacy. It can sometimes be fixed by convention: for example, a litigant in England can write a letter marked 'without prejudice' to another proposing a settlement, and this letter cannot be used in evidence. But most circumstances lack such clear and convenient rules, and the electronic nature of communication often means that 'just stepping outside for a minute' isn't an option. What then?

A related issue is anonymity. Until the industrial revolution, most people lived in small villages, and it was a relief – in fact a revolution – to move into a town. You could change your religion, or vote for a land-reform candidate, without your landlord throwing you off your farm. In a number of ways, the effect of the Internet has been to take us back to an 'electronic village': electronic communications have not only shrunk distance, but in some ways our freedom too.

Can technology help? To make things a bit more concrete, let's consider some people with specific privacy problems.

1. Andrew is a missionary in Texas whose website has attracted a number of converts in Iran. That country executes Muslim citizens who change their religion. He suspects that some of the people who've contacted him aren't real converts, but religious policemen hunting for apostates. He can't tell a policeman apart from a real convert. What sort of technology should he use to communicate privately with converts?

2. Bella is your ten-year-old daughter, who's been warned by her teacher to remain anonymous online. What sort of training should you give her?

3. Charles is a psychoanalyst who sees private patients suffering from depression, anxiety and other problems. Previously he practised in a nondescript house in town which his patients could visit discreetly. Since lockdown, he's had to use tools like Skype and Zoom. What's prudent practice to protect patient privacy?

4. Dai is a human-rights worker in Vietnam, in contact with people trying to set up independent trade unions, microfinance cooperatives and the like. The police harass her frequently. How should she communicate with colleagues?

5. Elizabeth works as an analyst for an investment bank that's advising on a merger. She wants ways of investigating a takeover target without letting the target get wind of her interest – or even learn that anybody at all is interested. Her opponents are people like her at other firms.

6. Firoz is a gay man who lives in Tehran, where being gay is a capital offence. He'd like some way to download porn and perhaps contact other gay men without getting hanged.

7. Graziano is a magistrate in Palermo setting up a hotline to let people tip off the authorities about Mafia activity. He knows that some of the cops who staff the office in future will be in the Mafia's pay – and that potential informants know this too. How does he limit the damage that corrupt cops can do?

8. Hristo helps refugees enter the UK so they can claim asylum. Most of his clients are fleeing wars or bad government in the Middle East and North Africa. He operates from Belgium and gets clients into trucks or on to speedboats depending on the weather. He needs to coordinate with colleagues in France, Britain and elsewhere. How can they do this despite surveillance from assorted security and intelligence agencies?

9. Irene is an investigative journalist on a combative newspaper who invites whistleblowers to contact her. She dreams of landing the next Ed Snowden. What preparations should she make in case she does get contacted by a major source that the government would try hard to unmask?

10. Justin is running for elected office. Irene would happily dig the dirt on his family; and there are many other people who want to read his email, send a racist tweet from his social media account, or wire his campaign war chest to North Korea. How can he frustrate them?

Privacy isn't just about encrypting messages. If Andrew tells his converts to download and use Wickr, then the police spies pretending to be converts will get the national firewall to detect anyone who uses it. Andrew has to make his traffic look innocuous – so that the police can't spot converts even when they know what apostate traffic looks like. If only a few dozen people use Wickr, the police can just break all their doors down. So it's not just about whether Wickr is a more secure product than Signal or Skype, but how many people in that country use it.

And while technical measures may solve part of Andrew's problem, they won't be much use with Bella's. One risk to children is that they'll say something careless that may embarrass them later. Another is that political and media scaremongering about child safety gets in the way of their welfare. Most of your effort will go into educating her. As Bella grows up, she'll have to become adept with the tools the rest of her peer group use; and soon enough she'll adopt her security procedures from them more than from you. You have to impart understanding, not rituals.

The intensity of attacks will vary widely. Andrew and Firoz might face only sporadic interest, while Graziano, Hristo and Justin have capable motivated opponents. As for Dai, she's frequently put under surveillance. She's not using anonymous communications to protect herself, but to protect others who haven't come to the police's attention yet.

There are radically different incentives. Andrew, Charles, Dai, Graziano and Irene go to some trouble to protect vulnerable people they deal with, while the sites in which Firoz is interested don't care much about his safety. Andrew, Dai, Graziano and Hristo all have to think about dishonest insiders. In Justin's case it's careless insiders: the juicy stuff that the Russians would like to give to Irene lives in the personal accounts of his campaign volunteers, as well as in the personal accounts of friends and family who're hard to include in any organised defensive effort.

There are different thresholds for success and failure. Hristo can only be jailed if the police prove a case against him beyond reasonable doubt; Irene can take down Justin if she can defend a libel suit on the balance of the evidence; while mere suspicion could be bad news for Elizabeth or Firoz. And there are different costs of failure: Elizabeth might lose some money if she screws up, while Justin could lose his career and Firoz could lose his life.

We discussed in section 22.2.1 how people who don't want their phone calls traced buy prepaid mobile phones, use them for a while, and throw them away. But these *burners*, as they're sometimes called, are hard to use properly; even Al-Qaida couldn't do it right. So what's the state of play for hard privacy online?

25.4.1 Anonymous messaging devices

As we discussed in section 2.2.1.10, investigators often get much of their information from traffic analysis. Regardless of whether people use email, a messaging service or the plain old telephone service, access to the social graph lets policemen map out friendship networks – and the marketers do this too when they

can get their hands on it [598]. In the old days, encrypting your email traffic could be dangerous; if you were one of only 20 people in the country using PGP, that made you a suspect. It's more complex now that most people use webmail services that are TLS encrypted by default, but the same principles apply.

People under government surveillance like Hristo learned that normal privacy apps like WhatsApp or Signal aren't enough on their own, even if lots of other people use them for innocuous purposes. Suppose Hristo uses Signal to arrange for Kevan to bring eight people across the English Channel in a speedboat. But if a Royal Navy cutter arrests Kevan and they find Hristo's messages on his phone, he faces extradition. If Kevan, or Hristo, also uses their phone to chat with their family, it might help the police to map their network using traffic analysis. There's not just an issue of making networks hard to trace, but about what evidence can be seized when people are caught. Similar problems are faced by Dai and by Graziano's undercover operatives.

So we've seen the development of a market for 'crypto phones' which not only provide encrypted messaging, but try to support *operational security* (Opsec) as well. We discussed Opsec in the corporate context in section 3.3.4, but it matters even more here. The first crypto phone on general sale was probably Silent Circle's Blackphone in 2014, which was sold to government agencies, special forces and human-rights workers. There have since been a number of competing systems. Ed Caesar describes some of the people who promoted crypto phone businesses out of the Cyberbunker in Germany, which was the country's biggest hoster of illegal web sites until it was raided and shut down in September 2019 [364]. The handsets are typically modified so you can't run apps (which could spy on you); they may have the microphone and camera disabled so GCHQ can't turn them into monitoring devices; GPS may also be disabled; they can't be read out by standard police forensic kiosks; and they're part of a closed system consisting of both phones and messaging servers where you don't identify the other party by a phone number but by a user ID. Crypto phone firms found that some people were prepared to pay over a thousand dollars or Euros for a handset, and the same again for a six-monthly subscription to the associated service. The market includes all sorts of people, from cryptocurrency operators and spies through money launderers to drug dealers. Network effects apply in covert communities too; Hristo, Kevan and the rest of the gang all need to use the same system. And as some people smugglers also smuggle drugs, and some smugglers make enough money to need fancy tax accountants who also work for the cryptocurrency crowd, network effects can drag in all sorts of people who seek privacy from state surveillance for reasons both good and bad.

The emerging pattern is that, thanks to network effects, one crypto phone system gets used ever more widely, until enough of its users are police targets and the authorities bust it. For the benefit of non-UK readers, I might mention here that newspapers of the left and right see Hristo and his human cargo in somewhat different terms. While some immigrant communities see Hristo's operation as a family reunification service, the conservative press stigmatises refugees and ministers have made immigration offences a higher priority for the agencies than organised acquisitive crime. So what can Hristo buy to keep GCHQ off his back?

Until 2016, the market leader was Ennetcom, a Dutch company which used

a private network of messaging servers in Canada to support anonymous user IDs. In April of that year, the Dutch and Canadian authorities raided them and arrested the owner, who had been involved with CyberBunker. In 2017, it was the turn of PGP Safe; four Dutchmen were arrested [1081]. The following year, the Dutch police also claimed to have broken a cryptophone system called Iron Chat [792]. In 2018, the market leader was a company called Phantom Secure; the US, Australian and the Canadian authorities closed down that system [1133]. Its CEO Vincent Ramos pleaded guilty to supplying the phones to drug dealers worldwide, and at his sentencing hearing, prosecutors read out a message he sent to colleagues: "We are f–ing rich man ... get the f–ing Range Rover brand new. Cuz I just closed a lot of business. This week man. Sinaloa cartel, that's what's up" [278]. He got nine years in jail. The next market leader, EncroChat, used modified Android phones. In 2020, the French and Dutch police hacked its main server and infected all 50,000 devices in use worldwide with law-enforcement malware that copied their messages in real time to the police. On June 13th, EncroChat realised they'd been hacked and advised their customers to get rid of their phones at once [1922]. Hundreds of arrests followed all over Europe [572].

So policemen like Graziano have a standard playbook for taking down crypto phone systems. But he may also use them to protect those of his sources who remain emplaced in the gangs and in the communities in which they swim. Indeed, when PGP first came out in the 1990s, it was adopted by the Provisional IRA in their insurgency against British rule in Northern Ireland. Up till then, a big headache for the police had been making unobtrusive regular contact with IRA informers, who lived in a nationalist community that hated the police and where informers were killed. PGP made contact easy. An informer simply had to tell his handler his private key, and the cops could collect all his traffic. He could even report in by sending an encrypted email to himself.

25.4.2 Social support

The journalist Irene probably has the hardest task. If she's approached by a senior civil servant who wants to spill the beans on the government's latest folly, then as soon as the story appears the 'mole hunt' will begin. Her informant – let's call her Liz – will now be hunted by the police and intelligence apparatus. How can Irene help Liz minimise the probability of being identified, fired, and prosecuted? We discussed whistleblowing briefly in section 2.3.6, where we saw that technical security is usually only one of the problems facing a whistleblower – and often not the most serious.

The big problem is establishing trust, and that is a two-sided process. Irene will need to assess Liz as a source. Does she have a real story to tell? Why's she telling it? Is it a semi-authorised leak, which she's offering with the tacit approval of her minister as part of a political game? Can her story be stood up with enough evidence, in case someone sues for libel? Is it a provocation, designed to discredit Irene or the newspaper she works for? Is Liz vulnerable, and in need of emotional support? When the story comes out, who else could have leaked it? If a hundred people could have leaked it, you can talk about anonymity; if the anonymity set size is only ten, you're talking more about plausible deniability, and Irene will want to talk to Liz about what happens

when the PM's goons interrogate her. But in many cases the whistleblower will be completely exposed once the story comes out. For example, if Liz's complaint is that a minister tried to rape her, then the conversation with Irene will be about getting support and about whether people will believe her, rather than about how to use Signal.

So best practice is for Irene to meet Liz in person as soon as possible after Liz makes contact. If Liz may be targeted by state actors, but has a reasonable chance of staying anonymous, Irene can give her a burner phone to establish a chain of contact that's independent of her existing home and work devices. If Liz is one of ten suspects after the story breaks and the Prime Minister starts shouting at the Director of the Security Service, then she'd better assume that all ten of them will have all their known devices compromised by tea-time.

When Ed Snowden decided to blow the whistle on illegal surveillance, he initially had difficulty getting a journalist to use PGP encryption. Afterwards, many newspapers rushed to provide technical means for whistleblowers to contact them, publishing PGP keys, the mobile numbers of journalists who use Signal, and a facility called SecureDrop that enables people to upload files. Mansoor Ahmed-Rengers, Darija Halatova, Ilia Shumailov and I did a study of such mechanisms and found they suffer from two types of problem [31]. First, such mechanisms are hard to use. We discussed in section 3.2.1 how security usability research started from the difficulty of using PGP, and the problem is still there. Second, a whistleblower needs to understand the hazards in order to devise sensible operational security procedures, but a typical newspaper doesn't discuss them the way that, for example, this chapter does. So Irene might want to give Liz not just a burner phone but a training session on how to use tools like Tails and Tor to upload files to SecureDrop[2]. (A crypto phone would be more usable but Irene probably doesn't have the budget, and if Liz were caught with one, it could be a giveaway.)

It would be a mistake, though, to think of Liz as Irene's typical source. Most whistleblowers are in an anonymity set of size one, and their disclosures are not about state secrets but about fraud and abuse. In section 12.2.6 we saw that whistleblowers stopped more of the really serious fraud than either auditors or regulators. But often a decision to expose wrongdoing may carry some personal cost, such as getting fired or being stigmatised. Social support is often the key. It was only after several women who'd been raped by Harvey Weinstein found the courage to speak out that dozens of others came forward too.

Support is critical for many of our other users, too. Charles the psychoanalyst knows that the privacy he can offer his patient, whom we might call Mary, is essential to the therapeutic work. The move from an office to videoconferencing not only creates some (small) actual risks but makes privacy less comprehensible to both, undermining its role as a facilitator in therapy. Mary might be afraid that if her employer discovers she's having therapy, she might be stigmatised by colleagues or passed over for promotion. In most cases the

[2]Even then, there's lots more a journalist ought to be aware of, such as the machine identification codes that modern printers embed in documents and which we discussed in section 24.4.3. They were used to trace Reality Winner, an NSA whistleblower who leaked an NSA document describing Russian interference in the 2016 election and got 63 months jail [174].

fear will be much greater than the actual risk, but sometimes the risk could be real: she might be Dai, or Irene, or Liz. So the therapeutic environment must calm her and inspire confidence. Charles cannot start off the relationship with a detailed briefing on opsec of the kind that Irene might give Liz at their first meeting. Privacy advice, if any, may have to be drip-fed along with the rest of the support.

When dealing with children such as Bella, the priority is also providing a calm and reassuring environment in which they can learn and develop. Sensible parents will see through the scaremongering around child safety; the rate at which children are abducted and murdered by strangers is about one in ten million population per year, a rate so low that rational people will ignore it. Your mission as a parent is to help your children grow into empowered citizens, not to train them to cower from imaginary monsters.

Dai, the activist, is also a giver of support, to the people she's trying to recruit. Her case is much more tricky than Charles's because the authorities are trying to stop her being effective. I assume she's known to the authorities and under intermittent surveillance.

Human-rights workers such as Dai do indeed use common tools such as Skype, Tails, Tor and PGP to protect their traffic, but the attacks to which they're subjected are not just technical; they're the stuff of spy novels. The police enter their homes covertly to implant rootkits that sniff passwords, and room bugs to listen to conversations. When they encrypt a phone call, they have to wonder whether the secret police are getting one side of it (or both) from a hidden microphone. Sometimes the microphone isn't all that hidden; we've heard from activists of the police standing openly outside their house pointing a shotgun mike at the window.

Countering such attacks requires tradecraft in turn. Some of this is just like in spy movies: leaving telltales to detect covert entry, keeping your laptop with you at all times, and holding sensitive conversations in places that are hard to bug. Other aspects of it are different: human-rights workers (like journalists but unlike spies) need to avoid breaking the law, and they also need to nurture multiple support structures – not just giving covert support to recruits downstream, but receiving overt support from overseas NGOs and governments. And to make recruits, they also need – while under intermittent observation – to make covert contact with people who aren't themselves under suspicion. Dai's case is the reverse of Charles's, as when she acquires a new recruit, training them in tradecraft is part of the induction, socialisation and support process.

If you want to learn about what works and what doesn't in tradecraft, then human-rights workers are the people to talk to. (The spies and smugglers may know more but they're not talking.) The emerging picture is that the behaviour of both police and nonviolent government opponents is embedded in how a society operates, and evolves over time. There's a complex game where all but the most totalitarian rulers attempt to marginalise, tame or co-opt their opponents, while the opposition movements evolve in response. Any movement that starts being too nice to an unpopular ruler will lose credibility and be displaced by others. The groups with the best opsec will be able to grow fastest, and the most militant groups may have the greatest credibility. Pushed too hard,

nonviolent opposition can spawn either open insurrection or violent terrorism (and rulers who denounce nonviolent opposition as 'terrorism' may invite just that). So a smart secret police chief will cut Dai some slack, watch what she gets up to, and play a long game; the Putin philosophy is to tolerate rebel movements until you can figure out how to lead them. Just in case things heat up later, he'll be sparing in the use of some of his capabilities, so he's got stuff in reserve for which she hasn't developed countermeasures.

25.4.3 Living off the land

Irene, Charles and Dai may find that their privacy tactics are influenced by the kind of support they have to give or receive, but they have something else in common – that they have to make the smartest use they can of what's available rather than buying or building special tools. We might perhaps call this *living off the land*[3].

In the old days, covertness could mean hiding in plain sight. Every country's elite has places to hang out, so if a senior civil servant wants to meet an eminent journalist, they can chat openly at a gentlemen's club in London or a country club in Virginia without anyone taking any notice. Such mechanisms allowed people to make contact discreetly, and establish trust at the same time.

So the first thing to ask when trying to improvise anonymous communications is what clubs or platforms you already share. One of the hard cases is China, which blocks most of the services familiar to us at the Great Firewall. Even there, we find platforms open to user content that have encrypted communications: two examples are Linkedin and Amazon book reviews. In the case of Iran, Andrew will have to figure out whether messaging systems such as Skype and Signal are sufficiently widely used there for their use not to be suspicious.

The second thing you have to think through is the threat model. One thing many of our users have in common is intermittent threat: most of the time there's no threat at all, but just occasionally it may become severe. Even a large secret police force can only work so many files at once. Most of the time, nobody's interested in Mary, or in Firoz either. However, if Mary suddenly becomes a celebrity, people will get interested in her mental health quickly enough. If the government suddenly decides to go after Nur, then Skype might provide her with cover in Iran, but not in Saudi Arabia – because Skype belongs to Microsoft which generally complies with government warrants except in rogue states. Even in Iran, some opsec is needed. If Andrew uses Skype to talk to Nur then he'd better not use the same username (or IP address) to talk to all his other converts too, or the religious police will learn it from their bogus convert and come knocking.

A third factor is capability, including support, and motivation. Of all our users, Elizabeth the investment banker may be the simplest case. Her work is lawful and she has an IT team for support. Tor provides fairly good anonymity if used with care, and the stakes are low; if a target suspects her interest,

[3]This phrase is also used of hackers who attack systems by exploiting the target's vulnerabilities directly as they need to, and don't leave remote access Trojans behind. It seems appropriate in this context too.

she only loses some money, not her life. Graziano faces higher risks but has an experienced police organisation at his back. Justin is also playing for high stakes, but has a much less tractable management problem. An election campaign is a long slog of fundraising with dozens of volunteers who're hard to discipline and whose focus is victory rather than security. Liz faces significant risks and the quality of support available from Irene may vary. Dai, Firoz, Hristo and Nur all face extreme hazards without any capable technical support.

Finally there's the problem of forensics. I'll discuss this in detail later in section 26.5, but the main problem for the police is the sheer volume of data found when searching a house nowadays: there can be terabytes of data scattered over laptops, phones, tablets, cameras, TVs, memory sticks and all sorts of other devices. If you don't want a needle to be found, build a larger haystack. So Firoz might have a lot of electronic junk scattered around his apartment, as cover for the memory stick that has the contraband stashed in an encrypted volume. And there are many ad-hoc ways in which content can be made inaccessible to the casual searcher; he might damage the memory stick in some repairable way, or just hide it physically. The same approach might be taken by Nur, or anyone for whom a police raid might be bad news.

This all comes back to tradecraft. What works will vary from one place and time to another, as it depends on what the local opponents actually do. To defeat routine traffic analysis, it might be enough to get a day job as a receptionist: if everyone in town calls the doctors' surgery, then the fact that someone called the surgery conveys little information.

25.4.4 Putting it all together

Returning now to our list of users, how can we sum up what we've learned?

1. The missionary, Andrew, has one of the hardest secure communication tasks. He can't meet his converts to train them in opsec, and needs to use something that's available and inconspicuous. Perhaps the simplest solution for him is to use Skype or WhatsApp.

2. In the case of your daughter Bella, the goal is to help her grow into a capable adult. I'd never dream of getting my grandkids to use Tor; that's just creepy. What I do is to talk about scams, phishing and other abuses from time to time round the dinner table. The kids enjoy this and slowly absorb the art of adversarial thinking. It's all in the same spirit as the board games we play.

3. The psychoanalyst, Charles, should have a basic awareness of the risks and the possible mitigations. As he gets to know his patient Mary he may occasionally make suggestions he thinks are relevant and needful, so long as they go with the flow and empower her rather than scaring her. But he may also be reluctant to make suggestions if this goes against the clinical method to which he is committed, by undermining her trust in the therapeutic environment. It may be too hard to negotiate this environment; informed consent is a difficult issue in therapy because of the asymmetric power relationship between the patient and the therapist.

In practice both parties may lack relevant knowledge, and even if Mary knows more about the risks than Charles, she may feel unable to offer any suggestions.

4. The human-rights activist Dai has one of the hardest jobs of all, but as she's shaken down by the secret police from time to time and works with other activists with whom she can share experiences, she can evolve good tradecraft over time.

5. The M&A analyst Elizabeth may well find that Tor does pretty well what she needs. Her main problem will be using it properly and paying attention to the kind of queries she makes of target websites so as not to give the game away.

6. Firoz is in a bad way, and quite frankly were I in his situation I'd set out on the walk to Germany. If that's not possible then he should not just use Tor, but get a Mac or Linux box so he's less exposed to porn-site malware. He'll need to think through in advance what happens if he gets raided by the police. (Perhaps he should join the Revolutionary Guard so the police won't raid him in the first place.)

7. Graziano also has a hard job. It's bad enough defending a covert network against one or two traitors at the client end (as Andrew must); defending against occasional treachery at the server side is even harder. Part of his solution might be a compartmented police record keeping system, as we described in section 10.2, to stop bent cops getting access to everything. He might also chat to informers using whatever mechanisms they use.

8. Hristo may see advantages in using a crypto phone, but when the cops crack it they may roll up his whole network. In his shoes I'd learn from Dai that in the long run the group with the best opsec wins out. So I'd focus on that, and educate my colleagues about traffic security. If we use a chat app such as Signal with ephemeral messages, and change phones and SIM cards regularly, then I can see which of my colleagues are disciplined, and decide who to trust with what.

9. Irene the journalist has one of the most challenging jobs of all. A journalist needs to be skilled not just at writing stories but at reading people, assessing truth and judging risk. An investigative journalist also needs tradecraft. Just as any journalist nowadays needs to know how to drive a search engine, a sleuth needs to know how to protect her sources. It's not enough to have some basic familiarity with privacy tech; she needs to know how to teach the right tactics to contacts who may be under extreme stress and at risk of their lives. That means understanding not just the people, but also the threats and the tools. (And just as this job becomes ever more critical and highly skilled, the budgets available to the press are collapsing, as Google and Facebook eat all their advertising.)

10. Justin also has a difficult problem. It's hard to protect short-lived high-consequence efforts staffed by enthusiastic volunteers who are hard to discipline and who may have unfixable bad technology habits. However he probably doesn't understand his vulnerability, and will just press on, hoping for the best.

Richard Clayton wrote a thesis on anonymity and traceability in cyberspace, which analysed how complicated network anonymity has become [442]. There are many ways in which even people who made no particular effort to hide themselves end up not being traceable. It's hard to establish responsibility when abusive traffic comes from a phone line in a multi-occupied student house, or a prepaid mobile phone. ISPs also often keep inadequate logs and can't trace traffic afterwards. But there are also many ways in which people who try to be anonymous, fail; eventually people make mistakes, regardless of how much effort they put into opsec. And technology is making opsec harder all the time. This even applies to government security and intelligence agencies.

25.4.5 The name's Bond. James Bond

We got a warning in January 2010 that traditional intelligence agency tradecraft, as described in the novels of Ian Fleming and John le Carré, was beginning to fray. The Israelis sent a team of 26 Mossad agents to Dubai to kill Mahmoud al-Mabhouh, a senior Hamas official who was there to buy arms from Iran. In the past such killings had been covert, but this time the UAE authorities collected and examined all the CCTV footage, correlating it with the agents' hotel stays and border crossings. It turned out twelve of them used British passports – many of them issued to Brits who'd emigrated to Israel, but with the agents' photos on them – along with six Irish, four French, three Australian and one German. Britain and Australia expelled Israeli diplomats for passport offences [307]. In the modern world of pervasive surveillance, biometrics at border controls and online passport databases make it a lot harder to travel under a false name.

A second warning came in 2013, when a report analysed the kidnapping of a Muslim cleric called Abu Omar in Italy in 2003, and pinned it on the CIA, leading to a number of agents being charged by the Italian police in absentia [1274]. The third warning came in 2014, when the Chinese stole the entire US security clearance database from the Office of Personnel Management, as I described in section 2.2.2; this included not just the entire US intelligence community but 22 million current and former federal employees. The weaponisation of personal information continues; the 2016 Investigatory Powers Act enabled the UK government to demand bulk personal datasets from firms who have them, giving the agencies access to credit records, medical records and much else. By the end of the decade, the military were worried that the Chinese were collecting personal information on every single enlisted person for use in future information warfare, while intelligence agencies were starting to wonder whether the age of traditional spying was over [1274]. The defence and intelligence communities have responded in various ways, with the Pentagon telling staff not to use consumer DNA testing kits and the Chinese apparently favouring more low-tech stuff like dead drops, but it's not clear that there's any silver bullet. It's hard to run covert operations when so much is known about almost everybody.

In this context, China's bid for 'AI supremacy' is concerning. The country's political structure encourages, rather than restrains, this technology's worst uses: President Xi wants an an all-seeing digital system of social control, patrolled by precog algorithms that identify potential dissenters in real time [48]. I discussed face recognition in section 17.3; as China's cities are straddled with

CCTV systems, they can surely follow people about. But how well will this work overall? The use of machine learning in multisensor data fusion applications isn't straightforward, and it tends not to work well or at all at social prediction – as we discussed earlier in this chapter. In section 26.4.1 we discuss how the Chinese system appears to be using the dissident Uighur population of Sinkiang as the test case, with substantial human-rights abuses which have led to US and EU sanctions against the Chinese firms involved.

Meanwhile, in our somewhat more chaotic democracies, it's hard to secure political campaigns from attack, as our discussion of Justin's case brings out. The resulting operational problems from the 2018 US election are discussed by Maciej Ceglowski [397], who also warns of the broader problems of securing elections. We turn to them next.

25.5 Elections

As I write in 2020, people are worried about the conduct and credibility of the forthcoming U.S, elections, following the controversy about Russian interference in 2016 in both the UK Brexit referendum and the US elections later that year. Because of the huge diversity of voting systems, US elections have for years been a testbed for voting technology. There have been very many attempts to defeat the will of the people, first by candidates, and more recently by external actors. We also have significant experience from the Commonwealth, which contains most of the other former British colonies; all of its member states hold elections of some form or another [329].

The story of election technology and its security is one of the co-evolution of attack and defence over centuries. In school, we all learned some variant of the history of how we evolved modern constitutions. Participatory government has long been widespread at the level of small groups such as villages, where everyone knows everyone else and decisions can be taken by consensus or by a majority; the problem is scaling it up to larger units such as cities and states. The Greeks and Romans experimented with mechanisms for selecting representatives to sit in assemblies, councils and courts but found that, all too often, democracy degenerated into oligarchy, or a monarch seized power. They devised constitutional mechanisms to reduce the risk of such failures, including the separation of powers, voting by geographical constituencies rather than by tribe, selecting officeholders by lot rather than by ballot, and term limits. Although the Roman Empire ended these experiments, the ideal persisted through papal elections and medieval guilds via Swiss and Italian city-states. In the English Civil War, a parliament seized power from the king and cut his head off; the settlement of 1689 made England a constitutional monarchy. The seventeenth century also saw the first assemblies in the New World, leading to the American revolution in the eighteenth century, where the Founding Fathers were inspired by the Greek and Roman model.

Behind it lies another story of how the elites who enjoyed power kept manipulating the system so as to hang on to it. Early elections had no privacy; Roman electors lined up behind their candidate, and voting by open outcry remained the norm until the nineteenth century, leading to bribery and intimidation. The

tension in England was about social class: barons acquired some rights in 1215, followed by other property-owners in a series of reforms. The first modern reform in 1832 introduced redistricting: few of the English cities that had sprung up in the industrial revolution had MPs, while other constituencies had few voters and the MP was selected by the local landowner. It took a whole series of reform bills to extend and equalise the franchise to men of successively lower wealth and income, but the high costs of campaigning limited political careers to the wealthy. Eventually, secret ballots were introduced in 1872. Meanwhile in America the story was more about race. The Civil War ended slavery and extended the franchise to all men; but after the failure of Reconstruction, former Confederate states devised literacy tests and other laws to stop black citizens voting. Only after World War I were women allowed to vote in either country. Abuses were rife: to this day, politicians in the UK, the USA and elsewhere try by fair means and foul to get their supporters to vote more than their opponents.

25.5.1 The history of voting machines

From the late 1800s there were waves of technological innovation that tried to push back on electoral abuses in America, a story told by Douglas Jones and Barbara Simons [991]. Many cities and states had political 'machines' that not only got out the vote but also manipulated it, exploiting the fact that elections in America are organised at state and county level rather than nationally as in Britain. In New York, Tammany Hall's Boss Tweed would sometimes stuff ballot boxes, and sometimes just have his precinct staff make up the results. To push back on this, inventors came up with everything from transparent ballot boxes to voting machines that clocked a mechanical counter when a lever was pulled.

Crooked politicians and officials adapted. In Louisiana, the Long brothers defeated the seals, set the count to the desired outcome and ran the state for years. Eventually people realised that the technicians in the county building who maintain and program the machines controlled the outcome. Mechanical voting machines had about 100 bits of programmability, typically in the form of cotter pins and other mechanical linkages, which nobody else understood. Wear and tear could also cover tampering; the technicians could cause an undercount for a candidate they didn't favour by knocking a few teeth off the relevant gearwheel.

25.5.2 Hanging chads

Inventors devised a competing type of machine that punched a hole in a paper roll, inspired by the player piano; once punched cards were popularised by tabulating machines and computers, they became widely used. The idea was that a vote punched as a hole in a card is both human-readable and capable of being counted quickly by machine. It's also anonymous once dropped into a ballot box (unless you worry about fingerprints).

In the 2000 US presidential election, the result turned on Florida, which used punched-card machines, and the recount involved arguing over chads – the little rectangles of cardboard that a voter punched out of the card. Was a 'hanging

chad', still attached to the card, a valid vote? What about a dimple, where the punch hadn't penetrated? Vote-counting machines rejected over 100,000 votes while George Bush's majority over Al Gore was only 537. Eventually the Supreme Court halted a recount, giving the election to Bush. This created such controversy that in 2002 Congress passed the Help America Vote Act (HAVA) which allocated $3.8 billion for the purchase of newer election equipment.

A gold rush followed as companies scrambled to build and sell machines into this huge new market. This alarmed security engineers. In fact, as the Florida recount was underway, I was at the Applications Security conference in New Orleans, whose attendees included many NSA and defense contractor staff, and we organised a debate. Even though politicians thought that mechanical or paper voting systems should be replaced with electronics as quickly as possible, security experts didn't agree. A large majority voted, on an old-fashioned show of hands, that we didn't trust electronic elections. A 1988 report by Roy Saltman at the National Bureau of Standards had already spelled out most of what was likely to go wrong [1641].

Some of the new products were *direct recording electronic* (DRE) machines, the descendants of the lever machines of the 19th century, which typically presented the candidates and other ballot options on a screen, then recorded the voter's input. Later research showed that about a quarter of votes made with a DRE machine contained at least one error – defined as a vote different from voter intent. Such 'vote flipping' was widely reported in Sarasota, Florida, in 2006, and it was unclear whether the root cause was usability or technology (depending for example on how you classify insensitive touch screens). Either way, a third of voters ignored wrong votes on the review screen [991].

Many problems were reported in the 2002 elections [806]; the following summer, the leading voting-machine supplier Diebold left its source code on an open web site in a security lapse. Yoshi Kohno and colleagues analysed it and found that the equipment was "far below even the minimal standards of security expected in other contexts": voters could cast unlimited votes, insiders could identify voters, and outsiders could also hack the system [1075]. Almost on cue, Diebold CEO Walden O'Dell, who was active in the campaign to re-elect President Bush, wrote 'I am committed to helping Ohio deliver its electoral votes to the President next year' [1987]. This led to uproar, and calls for a law to implement Yoshi's key recommendation, that there should be a voter-verifiable audit trail. (The voting researcher Rebecca Mercuri had argued as early as 1992 that DRE equipment should display the voter's choice on a paper roll behind a window and get them to validate it prior to casting [1295].) In some DRE machines this is provided in the form of a nonvolatile memory cartridge that records all voter actions, but this creates a tension with privacy. Other DRE machines had no audit trail at all; all an auditor could do was ask them to print out the same result again.

25.5.3 Optical scan

Most of the non-DRE equipment consisted of *optical-scan* machines that would scan a ballot paper or card that the voter had completed, whether with a pen or a special ballot-marking device, and then dropped into a ballot box. Optical

scan systems had been around since the 1980s, and had evolved from the mark-sense scanners used to score multiple-choice tests in schools.

In the following electoral cycle, Californian Secretary of State Debra Bowen authorized a large team of computer scientists, led by University of California professors David Wagner and Matt Bishop, to do a thorough evaluation of the state's voting systems. The reports made depressing reading [306]. All the DRE voting systems they examined had serious design flaws that led directly to specific vulnerabilities that attackers could exploit to affect election outcomes. All of the previously approved voting machines – by Diebold, Hart and Sequoia – had their certification withdrawn, and a late-submitted system from ES&S was also decertified. California could take such radical action, as perhaps three-quarters of the nine million people who voted in 2004 did so using a paper or optical-scan ballot.

A similar inspection of Florida equipment was carried out by scientists at Florida State University, who reported a bundle of new vulnerabilities in the Diebold equipment in July 2007 [749]. Ohio followed suit and came to similar conclusions. All the evaluated equipment had serious security failings: data that should have been encrypted wasn't; encryption done badly (for example, the key stored in the clear next to the ciphertext); buffer overflows; useless physical security; SQL injection; audit logs that could be tampered with; and undocumented back doors [1261].

But if you abandon DRE machines for optical scanning of paper ballots, as most US counties have since 2006, you can do a hand recount if a close result is challenged. But there are still lots of things to go wrong.

First, hundreds of counties use ballot-marking devices, so that the voter makes their choices on a touch screen, after which the machine prints out a voting form they can inspect visually and drop into a ballot box. But some machines make separate human-readable and machine-readable marks, and if such a machine can be hacked, it could print a ballot card where the text says 'Gore' but the barcode says 'Bush'. So there's a lot of detail around what you inspect, and how; best practice is to design for a *risk-limiting audit*. In the UK, the gold standard is still the hand-marked paper ballot, but in the USA the vendors of ballot-marking machines have enlisted disability rights campaigners to sell their equipment.

Our experience in the UK is broadly comparable, although we never adopted voting machines. Tony Blair's government progressively expanded the use of postal and other absentee forms of ballot, which was criticised by opposition parties as it made vote-buying and intimidation easier. Party workers (of which Blair's Labour party had more) could pressure voters into opting for a postal ballot, then collect their ballot forms, fill them out, and submit them. Plans to extend voting from the post to email and text were criticised for making this existing low-grade abuse easier and potentially open to automation. Finally, in the May 2007 local government elections, electronic voting pilots were held in eleven areas around the UK. Two of my postdocs acted as scrutineers in the Bedford election, and observed the same kind of shambles that had been reported at various US elections. The counting was slower than with paper; the system (optical-scan software) had a high error rate, resulting in many

more ballots than expected being sent to human adjudicators for decision. (The printers had changed the ink halfway through the print run, so half the ballot papers were 'the wrong shade of black'.) Even worse, the software sometimes sent the same ballot paper to multiple adjudicators, and it wasn't clear which of their decisions got counted. In the end, so that everyone could go home, the returning officer accepted a letter of assurance (written on the spot by the vendor) saying that no vote would have been miscounted as a result. Yet the exercise left the representatives from the various parties with serious misgivings. The Open Rights Group, which organised the volunteers, reported that it could not express confidence in the results for the areas observed [1472]. The Electoral Commission did not disagree, and this experience persuaded the UK to continue using hand-counted, hand-marked paper ballots to this day. (UK election abuses happen at other places in the kill chain, from voter registration through postal voting abuses to breaches of campaign finance limits: so fixing the computers won't be enough to fix the problems.)

25.5.4 Software independence

This experience brought home both the importance of, and the difficulty of achieving, *software independence* – the property that an undetected change or error in voting software cannot cause an undetectable change or error in an election outcome [1608]. We must assume that vote-counting software is buggy and it may be malicious, so we should not have to depend on it, and the possibility of a manual recount is a vital mitigation. But how do you do that in practice? In Bedford the candidates reckoned that a manual recount would have led to the same result but with a different majority, and didn't want to spend another 20 hours on a full manual recount.

The consensus view in 2020 is that systems must be designed to support a *risk-limiting audit* that can place strict bounds on the probability of fraud or error arising as a result of things going wrong with the software. For optical scan, this might mean keeping all the votes from each ballot box in a separate bundle, so that a candidate could challenge "let's do a hand count of boxes 17, 37 and 169" and this could be completed quickly. If the count is close, or discrepancies are found, you can hand-count more boxes. (In fact, an argument over partial versus state-wide recounts figured in the Bush v Gore lawsuit in 2000.)

Cryptographers have tried to make vote-tallying more verifiable. Research into cryptographic election mechanisms goes back to the early 1980s, when David Chaum proposed giving voters a digital ballot token constructed using the same general techniques as digital cash, which they can spend with the candidate of their choice. In section 5.7.7 I described the mechanism: it's an interesting crypto design problem as you need to support anonymity and auditability at the same time. The voter needs to be confident that their vote has been tallied properly but in order to prevent vote buying they must not be able to prove this to anybody else – the vote must be *receipt-free*.

After more than thirty years of research, there are now well-understood mechanisms for this. For example, the free Election Guard system from Josh Benaloh and colleagues at Microsoft Research allows digital ballots to be cast

in a vote collection device such as a scanner or ballot-marker in such a way that the encrypted ballots can be counted – the homomorphic property of El-Gamal encryption is used so that multiplying two encrypted votes has the same effect as adding two plaintext ones. A bit more work is required to ensure that all the ballots are well-formed and the result is decrypted properly, but the outcome is a software-independent count [223]. This was piloted in Fulton, Wisconsin, in 2020 in a primary election for Wisconsin Supreme Court candidates.

Cryptographic vote-tallying is marketed as 'end-to-end verifiable' but this claim is somewhat ambitious. It solves only the vote-tallying part of the problem. As with the electronic signature devices discussed in section 18.6.1, you don't have a trustworthy user interface, so you still have to worry about bugs and Trojans in the ballot-marking device or scanner. You still need the audit. You still have to worry about attacks on voter registration, on pollbooks, on result aggregation, and on the announcement of results. And if the vote collection device is an app on the voter's phone, you have to worry about vote-buying and intimidation, as with postal ballots. Then you also have to worry about phone malware, and about the quality of the design and implementation. A detailed evaluation of such an app that has been used in some US elections found dozens of problems [615].

25.5.5 Why electronic elections are hard

Another interesting threat emerged in the Netherlands. DRE voting machines had been introduced progressively during the 1990s, and cyber-rights activists were worried. They ran some tests and discovered that the machines from the leading vendor, Nedap, were vulnerable to a Tempest attack: using simple equipment, an observer sitting outside the polling station could see what party a voter had selected [785]. From the security engineer's perspective this was useful, as it led to the declassification by the German intelligence folk of a lot of Cold War Tempest material, as I discussed in section 19.3.2 (the Nedap machines are also used in Germany). The activists got the political result they wanted: the District Court in Amsterdam decertified all the Nedap machines.

As for other countries, the picture is mixed. In some elections in less-developed countries, the state has systematically censored opposition parties' websites and run denial-of-service attacks; in others (typically the most backward), elections are rigged by more traditional methods such as filing bogus criminal charges to get opposition candidates off the ballot, or just kidnapping and murdering them. The best survey of abuses worldwide may be the Commonwealth's 2020 report [329]. The news as I write this is of unrest following the election in Belarus where 'Europe's last dictator', Alexander Lukashenko, declared he'd won over 80% of the votes in an election at which exit polls suggested that his opponent Svetlana Tikhanovskaya had actually won 70% of the vote. His thugs compelled her to make a concession speech and drove her into exile in Lithuania, keeping her husband hostage. Lukashenko then put the resulting demonstrations down by force [611]. Another news story was the overthrow in a coup of the President of Mali, following allegations that he had stolen an election five months previously [1200].

In recent years there have also been many tussles over population registra-

tion; in section 7.4.2.2 I described how less developed countries rig elections by re-issuing the national ID card, and making cards harder to get for the ethnic groups less likely to support the president. Even where registration mechanisms are fairly robust, as in India with its Aadhaar biometric system mentioned in section 17.4, the authorities can attack voting rights directly: the government of Narendra Modi passed a law in 2019 to disenfranchise many Muslims, particularly those in border areas.

This is a very old playbook. As I already mentioned, right up until the twentieth century, electoral history in the UK was about whether poor people could vote, while in the USA it was about whether black people could vote. Even in Florida in 2000, more voters were disenfranchised as a result of registration abuses than there were ballots disputed because of hanging chads. And just as the government can bias an election by making it harder to vote if you haven't got a car, it could make it harder to vote if you haven't got a computer. There have also been lawsuits over whether the ballots, or the voting machines, were made so complex as to disenfranchise the less educated.

Several disputes over technical security have got to court. For example, the state of Georgia appears a complete mess as I write in 2020; after years of trying to make it harder to vote, failing to fix known flaws in Diebold machines and being targeted by the Russians, the state government was ordered by a court to replace its systems. The new systems were in meltdown during the June 2020 primaries, with insufficient capacity to meet voter demand [851].

However the main focus of attention has shifted to the use of social media in elections. Barack Obama used Facebook effectively in 2008 and 2012, prompting others to study social media; the 2016 election went to Donald Trump, who was not only much more skilful than Hilary Clinton at using Twitter, but ended up paying significantly less for his Facebook ads. As I explained in section 8.5, the ad auction mechanisms used by Google and Facebook multiply the amount that you bid by a factor called 'ad quality', which is the probability that people will click on the ad and, in the case of social media, share it. The outcome is extremism bias: inflammatory ads are cheaper.

Another factor in 2016 was Russian interference, as I described in section 2.2.3. Russian agents not only campaigned for Trump, running troll farms and social-media advertising campaigns aimed at suppressing black votes among other things; they hacked the Gmail of Clinton campaign chair John Podesta. They hacked into systems in Illinois and Florida (and probably some other states) and could have manipulated voter registration, but they opted not to pull the trigger on those attacks because they didn't need them; they hacked the electorate instead. Had Clinton won, then if either of those states had voted for her, evidence of 'fraud' could have emerged to undermine her presidency.

How will this all affect the election due in November 2020? As this book is due for release then, I will merely note that there's already been a fiasco over result aggregation in the Democratic primary in Iowa [637], and the Russians are once more running inflammatory pro-Republican campaigns online [1619]. Both Twitter and Facebook have removed postings by Trump and his associates containing false information about Covid [1030], and there is concern that he or others might use online media to undermine the electoral process, or confidence

in the results. Trump prepared the ground at the Republican National Convention by claiming he could only lose if the election was stolen. There is anxiety within Facebook that although Zuckerberg has said he'll block attempts at voter suppression, he's been giving the right wing an easier ride [1740]. In August, the major tech companies announced an alliance to fight election manipulation [963]. But what about a dispute over the result afterwards? There's over a century of American political history to warn us against looking for technological solutions to political problems.

In the different political culture of Europe, we have a long tradition of campaign finance limits (America did too before the Citizens' United decision of the Supreme Court turned it into a free-for-all). Parties can spend only so much per campaign, and per candidate; and most European countries forbid paid TV ads during campaigns. But enforcement has been getting steadily weaker. During the Brexit referendum, for example, both Leave campaigns exceeded the spending limit but just paid the £20,000 maximum fine. The Russian involvement in Brexit was largely in the form of financial contributions and further campaigning on social media. What might be done to block such abuses?

At the 2019 conference of the Open Rights Group, I argued that we should extend the advertising ban from TV ads to all ads on Facebook, Twitter, and YouTube. This is not just a matter of avoiding the worst of big-money politics of the USA, but also because political ads that are targeted at individuals rather than at everybody foster extremism and fragment political discourse. The politicians' job is to mediate conflicts between different stakeholders in society; if these groups end up in their own filter bubbles, then our politicians can be tempted to inflame conflicts instead. Banning ads will not be a panacea (India banned Facebook ads in 2019) but it will keep election contests more within the cultural and economic space with which Europeans are familiar.

Elections remain one of the tough security engineering problems. While the individual problems – such as voter registration, vote casting, vote counting, result aggregation and audit – all have reasonably robust solutions, putting them together into a robust system is nontrivial. Computer systems for registering electors, recording votes and tallying them have a number of properties which make them almost a pathological case for robust design, implementation, testing and deployment. First, the election date is immovable and, ready or not, the software must be deployed then. Second, different regions and countries have different requirements and they change over time. Third, in the long gap between elections, staff with experience move on and know-how is lost. Fourth, operating systems and other software must be updated to fix known vulnerabilities, and updates can also break security in unforeseen ways; a Windows update caused the EV2000 voting machine to highlight the last voter's choice to the next voter [991]. Yet most voting machines in use in the USA are no longer manufactured, so where are the updates to come from and how will they be tested? Finally, elections are high-stress events, which increases the likelihood of mistakes [1357].

Let's now look up from the engineering to the politics. In the event of attack, the winners don't want to investigate what might have gone wrong if they can possibly avoid it – as we saw in both the USA and the UK in 2016[4]. The

[4]As I write, litigation continues in an attempt to force the release of the redacted parts of

'customer' for an election is the losing side, and in the absence of any hope of redress – whether through the courts, or through the ballot box next time – trust in democracy's mechanisms can start to fail. But there is no 'designer' to ensure that the mechanisms and laws align all the way along the electoral cycle. On the contrary, it's typically the incumbent who tweaks the laws, buys the voting machines, and creates as many advantages for their own side, small and large, as the local political culture will tolerate. And while voting mechanisms can support a democratic consensus, they cannot replace it: there are too many other ways to undermine the outcome. If the underlying social contract erodes, a hyper-partisan environment can lead incumbents to feel they do not dare to cede power. In the worst cases the outcome can be civil war and failed states.

25.6 Summary

Some of the most challenging security engineering problems in 2020 have to do with the fact that as software becomes pervasive in the services we use and the devices around us, the design of these services and devices comes up against the underlying complexity in human societies. We looked at four examples. Self-driving cars can cope with empty desert roads but find real traffic with human drivers very much harder. Machine-learning mechanisms can go only so far; they may be brilliant at pattern matching but lack understanding, which opens up new possibilities of abuse at all levels in the stack – especially as people rush to use them for social prediction tasks for which they are intrinsically unsuited. Privacy-enhancing tools and techniques are one way to explore the security consequences of human complexity, but however hard we work to encrypt and anonymise things, social structure tends to show through one way or another. And finally, we have elections; when incumbent rulers are prepared to do everything they think they can get away with – whether within or beyond the law – to stay in office, we can learn a lot about the limits of both technology and law.

As more and more of human life moves online, so the criticality and the complexity of online applications grow at the same time. Many of the familiar problems come back again and again, in ever less tractable forms. Traditional software engineering tools helped developers get ever further up the mountain of systems complexity before they fell off. What sort of tools, techniques, and governance processes are appropriate for dealing with the complexity of real societies? And how does this interact with politics? These are the topics we will try to tackle in the third part of this book.

Research Problems

One bundle of research problems is around how to split responsibility between people and automation. HCI guru Ben Shneiderman argues that human control plus extensive automation is the sweet spot for systems to be reliable, safe and trustworthy [1723]. This is natural for flight-control systems and life-support

the Mueller report into that election.

machinery, but scaling it up to things like recommender systems and hate-speech detection is not trivial. How can humans do quality control on millions of filtering decisions being taken every second by a large tech company? And what should the governance on top of that look like? Underlying it all is a long debate about whether automation (including ML) is heading towards artificial intelligence or intelligence augmentation [87].

As automation involving ML becomes more pervasive, the questions may become broader. Architects and city planners will have to wrestle with how we design living and working environments that have to take into account the interests of multiple stakeholders. Then there will be global social and political questions around the coevolution of mechanisms and societies. In the second edition I said that "one of the critical research problems between now and the third edition of this book ... will be how protection mechanisms scale... how one goes about evolving 'security' (or any other emergent property) in a socio-technical system with billions of users." I noted that simple systems of rules, such as the multilevel security beloved of governments, were never natural, and people always had to break them to get their work done. So what lies beyond? We have more experience of that now; several large tech firms run systems with over a billion active users, and hundreds of firms have over a hundred million.

In such systems, the technology and the behaviour adapt to each other, but the system developers are much more powerful and have different incentives (they want data while users want privacy). The basic mechanisms that humans have for rule negotiation at scale are competition in the market and government regulation. Neither of these is adequate on its own, and the interaction between tech and politics may even undermine the machinery of selecting a government.

We engineers need to care about these issues, and try to understand them. In the third section of this book we'll try to tackle the broader policy and management questions (such as surveillance versus privacy), how the evolution of large complex systems can be managed and governed, and how technology can be regulated to meet social goals such as safety and privacy.

Further Reading

For an introduction to car security, you might first look at Charlie Miller and Chris Valasek's account of how they hacked a Jeep [1316], then at Craig Smith's 'Car Hackers' Handbook' if you want to dive into the technical detail [1792].

Nicolas Papernot's "Marauder's Map" may be the best introduction right now to the fast-moving field of adversarial machine learning [1493], while Gary McGraw and colleagues offer design principles plus a list of things to think about when working on the security of systems with machine-learning components [1267]. Google's Jeff Dean, its SVP of machine learning, describes the company's research on AI fairness at [528]. My own philosophical position on the AI versus IA debate may be found at [87].

As for personal privacy in the face of hostile state actors, that's a moving conflict as the tools evolve on both sides. One starting point might be the "Surveillance Self-Defence" page on EFF website [618]. There's an interesting

account by Ben Collier of the organisational and social dynamics of the Tor project, which maintains the leading online anonymity service, at [458]. For more technical depth, see section 20.4 on Tor, or the anonymity bibliography at [125].

The history of US voting systems is told by Douglas Jones and Barbara Simons [991]. The US National Academies of Sciences, Engineering and Medicine produced an extensive report on election security in response to the events of 2016 [1388]. More recently, the Commonwealth produced a guide to the electronic security of elections based on its own member states' very diverse experiences, also covering the whole cycle from registration through vote casting, tallying and communication of the results [329].