CHAPTER 2

DO NO HARM?

As colonial rulers of India, the British grew concerned about the abundance of cobras in Delhi. Governors therefore proposed a simple economic remedy: a bounty for cobra hides. The policy was a hit; so much so, that enterprising Indians started breeding cobras just for the bounty. Seeing a suspicious uptick in bounties paid, the British eventually cancelled the scheme. Rather than keep the now worthless snakes, breeders chose to loose the surplus serpents, causing the wild cobra population to surge past its previous levels, and defeating the point of the programme.[1]

Unintended consequences and externalities

Even the most benign, well-intended acts can have unexpected impacts. The 'cobra effect' would be no surprise to the French cultural theorist Paul Virilio.

> When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash; and when you invent electricity, you invent electrocution... Every technology carries its own negativity, which is invented at the same time as technical progress.[2]

For Virilio, technology's every yin has a corresponding yang, a range of unintended consequences birthed when the technology fails,

succeeds beyond expectations, or is simply used in unexpected ways. Philosopher Don Ihde argues that technologies have no fixed identities or meanings, and instead are *multistable*: people put tech to all sorts of uses beyond those the designer intended.[3] GPS was originally devised for the military, but since being released to civilians, GPS has spawned thousands of products and services, each with their own consequences. Satnavs have killed the road atlas and clogged village roads unwisely offered as shortcuts. Person-tracking software has both enhanced and eroded personal trust, saving lost children but ruining marriages and surprise parties alike. According to the *law of unintended consequences*, there will always be outcomes we overlook, but unintended does not mean unforeseeable. We can – and must – try to anticipate and mitigate the worst potential consequences.

A cousin of the unintended consequence is the *externality*. An externality is the economist's label for Someone Else's Problem, an effect that falls on someone outside the system. Passive smokers don't choose to smoke; instead they are victims of a negative externality, harmed by someone else's habit. Externalities can also be positive: one upside of public transport is that fewer pedestrians are killed by drunk drivers.

Unintended consequences affect familiar people in unknown ways, while externalities happen to people we've ignored. In other words, we overlook unintended consequences by not looking deeply enough, but we miss externalities because we were looking in the wrong places.

Externalities have been a sticky problem throughout the history of industry. A selfish, short-term focus has tempted many companies to harm their ecologies and futures. There's evidence, for example, that Exxon knew of CO_2's potential climate threat in 1977, but kept it quiet, preferring that society pay the cost.[4] Externalities also arise as a side effect of user-centred design. Focusing on fulfilling the goals and dreams of an individual user has caused tech companies to overlook impacts on non-users and wider society.[5] Airbnb is a dream for hosts and renters, but piles negative externalities onto the neighbourhood:

> At least in the short term, [Airbnb] reduces stock available for long-term renting or purchase […] Even then, though, a second externality remains: the impact on neighbors. Living next door to a permanent resident is very different than living next door to a constantly changing set of visitors that have no reason to invest in relationships, the neighborhood, or even a good night's sleep. To put it another way, small wonder hosts and guests love Airbnb: all of the costs are passed off to the folks who aren't earning a dime. —Ben Thompson[6]

The best way to quash externalities is, of course, to internalise them. Economists, as is their habit, typically suggest we do this with taxes or penalties. Many governments respond to environmental externalities with a *polluter-pays principle*, loading the cost onto the responsible party and nullifying the externality. Alternatively, they may choose to subsidise positive externalities, such as funding cycle-to-work schemes that also increase public fitness. If Airbnb chose to prioritise the neighbourhood's wellbeing – whether under consumer pressure, threat of fines, or as a result of some pang of social conscience – the externality would vanish. The community would become Airbnb's problem and neighbourhood-friendly policies would quickly follow.

Resolving externalities means we first have to recognise them, but often they lie in the shadows, falling on ignored minorities or existing only in a hazy future.

> If somebody robs a store, it's a crime and the state is all set and ready to nab the criminal. But if somebody steals from the commons and from the future, it's seen as entrepreneurial activity and the state cheers and gives them tax concessions rather than arresting them. We badly need an expanded concept of justice and fairness that takes mortgaging the future into account. —Ursula Franklin[7]

Algorithmic bias

Algorithmic bias – when supposedly impartial algorithms encode implicit prejudice – is a textbook example of unintended consequences. Bias has become one of tech's most notorious ethical issues, evidenced by several ugly examples: predictive policing software that

deems black people a higher reoffending risk than white people; YouTube's recommender system continually dragging people towards extreme content; networks that show high-paying job ads to men but not women.

Biased algorithms are clearly most dangerous when they rule over critical systems like justice or employment, but even a skewed commercial algorithm can have insidious effects. Individuals and groups can fall victim to *redlining*, denied products and services by biased software. The label comes originally from banking in the 1930s, when lenders drew on the city map to demarcate neighbourhoods (mostly home to black residents) they wouldn't lend to. Redlining today is less calculated but can be similarly damaging. Bloomberg found Amazon's same-day delivery service ignored majority-black neighbourhoods, such as Roxbury in Boston, despite all surrounding suburbs being eligible.[8] Even seemingly minor bias can stack up. Denied fast delivery, Roxbury residents may have to waste time and money buying from more expensive outlets: another brick in the wall of inequality.

None of these outcomes were planned; instead, they lay outside the scope of what technologists considered. No one looked deeply enough at the potential impacts on users, and no one thought to speak up for those who may be mistreated. This is still an industry failing. The public generally can't defend themselves against algorithmic bias or seek recourse. With no human in the loop, the decision rests in the hands of omniscient, unquestionable algorithmic gods. If your algorithmic luck is out, there's not much you can do except pray.

Sources of bias

AI ethicist Joanna Bryson claims algorithmic bias has three primary causes.[9] The first is poor training data. Data that's incomplete, unrepresentative, or improperly cleaned will always cause algorithmic blind spots. A facial recognition system trained only on white faces is guaranteed to be racist. This isn't just inconvenient, it's degrading: failing to recognise a face is failing to recognise someone's humanity.

Bias caused by patchy data typically hurts the underprivileged most. Rich people, with plenty of access to technology and detailed

financial histories, cast large data shadows; the poor or marginalised usually don't. Although an extensive data profile can sometimes be a risk to individuals, systemic 'data poverty' causes creeping harms to whole communities. Subpopulations become algorithmically invisible and are therefore unfairly treated; oppression is digitally re-enacted and amplified.

Bryson's second source of algorithmic bias is intentional prejudice. Algorithms offer an appealing way to launder bias beneath the illusion of objectivity, and many bigots within our own companies and governments have the power to twist algorithms towards their preferred intolerance. Intentional prejudice is always unethical but often legal. Different countries and states diverge strongly in attitudes and laws; today, it's legal in Kansas to fire someone for being gay, but not so in neighbouring Colorado. Prejudice can also come from outside the team. Microsoft's notorious chatbot Tay was programmed to learn through Twitter interaction, leaving it vulnerable to manipulation. Trolls leapt at the chance to goad Tay into making outrageous remarks and, when word of early successes spread, abuse quickly spiralled.

This hints at the third, most fundamental, source of bias: even the most complete dataset is suffused with human prejudice. Since Verbeek's mediation theory tells us we shouldn't separate technological and human action, technologies will mirror social biases by default. These biases run deep. Bryson and two colleagues trained a basic machine-learning system on a standard corpus of text and found 'every linguistic bias documented in psychology that [they had] looked for.'[10] According to Bryson, word embeddings – essentially, mathematical mappings of language – 'seem to know that insects are icky and flowers are beautiful' simply because those types of word are frequently paired. No surprise, then, that sentiment-analysis algorithms have inherited prejudice, deeming European names (Paul, Ellen) more pleasant than African-American names (Malik, Sheeren)[11] and ranking the word 'gay' as negative.[12] Even amid changing public opinion, this bias will only drain out slowly. Data always looks backward, meaning historical prejudice is frozen into a training corpus.

Inequity goes beyond language, of course: almost any data can be imbued with implicit bias. Even if it's illegal to consider someone's

race when calculating their credit score, every credit broker looks at the applicant's address, which is strongly correlated with race. Critics of the COMPAS crime prediction algorithm said it lent unfair weight to previous arrests and convictions. As law professor Ifeoma Ajunwa pointed out, 'If you're looking at how many convictions a person has and taking that as a neutral variable — well, that's not a neutral variable.'[13] It's well known that some populations are overpoliced and that ethnicity influences sentencing; these effects trickle into algorithmic logic. Even apparently innocuous decisions like where to build a distribution centre – presumably the root of Amazon's delivery bias – will depend on the local market, transport options, land values, and other factors heavy with implicit bias. Discrimination is already interwoven in the fabric of our tools and datasets.

Moral distribution

If bias is unintentional, is it really our problem? Surely it's not technology's job to fix every flaw in the human psyche? Again, remember mediation theory and its elegant dance of people and technology. At the scene of a crash, investigators will rightly ask whether the cyclist swerved, or whether the driver was fiddling with the stereo, but they may also check the car's brakes and ask who performed its last service. Technologies tend to spread moral responsibility between many actors.

As yet, we don't blame the technology itself, but if technology changes how users interpret the world, it follows that technologists influence people's moral choices. When things go wrong, the user and the technologist may both be to blame. Of course, technology is now a team sport: the era of solo hackers is long gone. Modern tech is made by teams of engineers, designers, product managers, and data scientists, relying on multiple underlying layers: AIs in apps on platforms, using shared libraries, plugged into various operating systems and protocols. Even if different teams or organisations have made each layer, all can be morally implicated. A team is only as trustworthy as its sleaziest partner. Build on a vulnerable platform, lock yourself into a disreputable social network, or share data with an abusive advertiser, and you will rightly bear some blame.

That's not to say we should blame technologists for every unin-

tended consequence: given tech's multistable nature, there will always be some outcomes that couldn't have been predicted. It seems unfair, say, to blame the architects of GPS for rural traffic jams. However, we had advanced warning of the bias issue. Even in the 1980s it was a documented problem, after doctors at St George's Hospital Medical School discovered their admissions algorithm, trained on the previous decade's decisions, was discriminatory:

> The computer used [implied] information to generate a score which was used to decide which applicants should be interviewed. Women and those from racial minorities had a reduced chance of being interviewed independent of academic considerations. —Stella Lowry and Gordon Macpherson[14]

Algorithmic bias may not be intentional, but it is negligent. Technologists might not have seen it coming, but we didn't much care to look. Convinced that technology is neutral and objective, we mistakenly assumed bias was impossible; concerned only with the productivity of the primary user, we overlooked impacts on wider society.

In the end, blame may not matter. Causal responsibility and moral responsibility don't always coincide; it's perhaps more useful to ask who has the power to fix things. Even if technologists didn't directly cause an ethical mishap, we still have a duty to try to resolve it. Regardless of intent, we must try to reduce harmful bias for the good of society and, secondarily, for the sake of our own reputations.

Moral relativism

Here we hit a classic ethical problem: doing the right thing sounds appealing, but what is right? What makes an algorithm fair? This quickly gets political. Should we aim for equal treatment or equal outcome? Treat everyone the same and you do nothing to address the systemic issues that perpetuate inequality. But pushing instead for more equal outcomes leads to accusations of meddling and reverse discrimination. Should algorithms simply reflect today's society or help us achieve a fairer world? Who chooses? Come to think it, is anyone's moral point of view more valid than anyone else's?

This is the seductive territory of *moral relativism*. A relativist argues

there's no one moral truth, no lone guiding star for behaviour; instead, ethical rules depend on social differences and vary across cultures.

Relativism usually stems from the well-meaning principles of tolerance and diversity. Holding all people accountable to the narrow values of a dominant culture – in other words, appointing a single party, country, or religion as the custodian of moral truth – has historically proved murderous, and relativists point out that people's individual beliefs are shaped by upbringing and evolve with experience. But conservatives typically decry moral relativism as postmodern flimsiness taken to dangerous extremes. Traditionalist philosopher Roger Scruton claims, 'A writer who says there are no truths, or that all truth is "merely relative", is asking you not to believe him. So don't.'

Globalisation is particularly challenging for relativism. Should we accept another society's choices that we find repellent? Should we do business with countries that actively discriminate, or where corruption is widespread? Moral relativism suggests a free-for-all: who are we to argue with the norms of another culture?

The philosophical debate rolls on, but for our practical purposes relativism is a dead end. If people can wriggle out of moral judgment by claiming their actions are culturally acceptable, morality itself becomes a questionable concept. Ad absurdum, if goodness is in the eye of the beholder, slave owners get to decide whether slavery is ethical. To make any kind of moral progress we need to be able to draw a line between acceptable and unacceptable behaviour. Fortunately, most cultures do agree on major rights and wrongs, such as murder and adultery. Forty-eight nations found enough common ground to encode basic moral principles into the Universal Declaration of Human Rights.

If we reject relativism, we have to reject it in the workplace too. Hardline tycoons might claim there's no room for personal morality in commerce: nice guys finish last. But if different cultures don't get to prescribe their own distinct ethical rules, nor does the business world. It's true that we all play many roles in life, and adapt our behaviours accordingly – a bruising tackle is acceptable on a football pitch, but not a wedding – but these roles are still underpinned by a common moral foundation, one that can't be substituted or switched

off at will. Morality doesn't stop at the front door of the office; business is, after all, made of people.

The technocracy trap

If we're to make moral progress, someone has to define what our ethical standards should be. Ideally, that's a matter for society itself. Elected officials make laws; citizens slowly develop social conventions. But disruptive technologies often burst onto the scene without warning, before these social or legal norms can emerge. By default, new technologies bear the ethical fingerprints of their creators, not of wider society. Given how heavily technology shapes modern culture, this means technologists have significant influence over social norms: ethical decisions that should be democratic are instead technocratic.

This should worry us. No single group should have a greater claim over the future than the public, and technologists don't have the diversity and worldly wisdom to be natural ethical authorities. Governments are belatedly creeping into action, and the public are now paying more attention to their technological lives, but the industry must also become more responsible. We must show that we deserve society's trust by engaging the public in the moral decisions that surround technology, and prioritising the good of all, not just our revenue streams.

Defining fairness

What sort of moral lines can we draw on algorithmic bias? First, we need to be precise. The word 'bias' is useful to a point: it's broadly understood and admirably simple, but we soon need more detail. Bias is an umbrella term for several types of imbalance: are we talking about sampling bias, innate structural bias, or explicit prejudice? To be specific, we must also be bold and discard some perfectly viable definitions.

Imagine you work for Tinder. If you want to ensure your matching algorithms are racially fair, what would fairness actually mean? Perhaps fairness is about exposure, meaning we should show users potential matches that reflect the local racial mix. If 30% of people in the user's city are black, we could tweak the algorithm so 30% of a

user's potential matches are black too. This sounds reasonable, but has an ugly flaw. The worlds of dating, sex, and love are riddled with human bias, meaning many people tend to stick to their own race when choosing a partner. So unless Tinder's users are bias-free, some people will be shown frequently to users who don't want to date someone of that race. We may achieve fair exposure, but certain races will be matched less often. One form of fairness forces another unfairness to the surface.

Should we aim instead for fair matching? Would it be better to declare that, whatever your race, you should have an equal chance of finding a partner through Tinder? This has the reverse problem. On today's online dating platforms, heterosexual white men give lower ratings to a woman if she is black.[15] To prioritise fair matching, the algorithm should actually *restrict* the racial mix by, for example, only showing black women to black men, who are more likely to respond positively. However, this is surely the opposite of racial fairness.

There's no way to square this circle. Thanks to the human biases surrounding dating, fair exposure can't be reconciled with fair matching. So perhaps we have to look elsewhere. Maybe a fair Tinder is one where people feel equally valued, or have similar levels of satisfaction with the service. This suggests we should care more about app usage patterns and satisfaction scores than the crude primary metrics of exposure and matching.

Any definition of fairness will be unfair from a different perspective. For some people, a fair algorithm is one that reflects today's society. For others, a fair algorithm must be an agent of social change. Some form of bias is logically, politically, and mathematically inevitable; nevertheless, someone has to make the call. Our decisions are the stars by which our algorithms will navigate. We must choose intelligently, considering the potential consequences and externalities of our choices.

Mitigating bias

Kate Crawford, NYU professor and co-founder of the AI Now Research Institute, is a leading expert in algorithmic bias. Crawford suggests teams invest in *fairness forensics* to mitigate bias. The first forensic step is simple: test your algorithms with a wide set of people

and solid benchmark data to spot problems before they occur. However, some benchmarks are themselves imperfect: common open-source face databases have historically skewed white and male. So teams may also want to screen their training and testing data itself for potential bias. Google's Facets software, for example, helps teams probe datasets for unexpected gaps or skews.

If these tests find bias, the simplest debiasing strategy is to improve the training data. A machine-learning algorithm trained on patchy data will always struggle, but simply throwing more data points at the problem often won't work. If 500,000 points of training data contain implicit bias, 5,000,000 data points probably will too; more active intervention may be needed.

Startup Gfycat found its facial recognition software frequently misidentified Asian people. The team resorted to what amounts to an 'Asian mode', extra code invoked if the system believed the subject had Asian facial features. While improved accuracy probably justifies Gfycat's hack, this sort of solution – the algorithm warning itself it's about to be racist, like some woke Clippy – isn't exactly scalable. It's exhausting and inefficient to play whack-a-mole with each new discrimination you discover, and cramming people into categories ('Is this person Asian? Is this person female?') to spark code branching also feels somewhat distasteful. Society sees classifiers like race and, increasingly, gender as spectrums rather than discrete buckets; our algorithms should too.

An even more forceful way to debias algorithms is to explicitly overrule them, gouging biased data or offensive associations from the system. Google Photos fixed their notorious 'gorilla' blunder – when the software classified a group of black friends as such – by overruling the algorithm: the team simply tore the word off the list of possible categories. In this case, the price of diminished classification power was clearly worth paying. Google has also added stop lists to search, after researchers found some racial terms generated appalling auto-complete suggestions.[16] Potentially problematic search stems now get no suggestions at all.

This sort of direct interference shouldn't be performed lightly. It solves only the one visible case, and the idea of forcing an algorithm to toe the desired line will spark accusations that technologists are imposing their personal politics on the world. Vetoes are best saved

for situations where the output is so clearly harmful that it demands an immediate fix.

Since bias can never be fully eliminated, at some point we face another tough decision: is the algorithm fair enough to be used? Is it ethically permissible to knowingly release a biased algorithm? The answer will depend in part on your preferred ethical framework; we'll discuss these shortly. A human decision will sometimes be preferable to a skewed algorithm: the more serious the implications of bias, the stronger the case for human involvement. But we shouldn't assume humans will always be more just. Like algorithms, humans are products of their cultures and environments, and can be alarmingly biased. Parole boards, for instance, are more likely to free convicts if the judges have just eaten.[17] After doing everything possible to ensure fairness, we might deem a lingering bias small enough to tolerate. In these systems we might release the system with caveats or interface controls to help users handle the bias, such as adding pronoun controls ('him/her/they') to translation software, allowing the user to override bias when translating from genderless languages.

While Crawford extols these forensic approaches, she also points out their shared weakness: they're only technical solutions. To truly address implicit bias we must consider it a human problem as well as a technical one. This means bringing bias into plain sight. Some academics choose to explicitly list their potential biases – a process known as *bracketing* – before starting a piece of research, and take note whenever they sense bias could be influencing their work. By exposing these biases, whether they stem from personal experience, previous findings, or pet theories, the researchers hope to approach their work with clearer minds and avoid drawing faulty conclusions. In tech, we could appropriate this idea, listing the ways in which our algorithms and data could demonstrate bias, then reviewing the algorithm's performance against this checklist.

Moral imagination

We can also pull bias out by the roots by getting better at spotting and addressing unintended consequences and externalities. For this, we need *moral imagination*: the ability to dream up and morally assess a range of future scenarios. Humans learn to use moral imagination

throughout their lives – indeed, we're the only species that can – but it isn't always easy to imagine the real impacts of technology. Our work is used asynchronously across the globe; we can never directly see the joy or pain we cause others. Fortunately, moral imagination can be trained. Morality isn't a genetic godsend; it's a muscle that needs exercise.

We can kick-start moral imagination with some straightforward prompts. The question 'What might happen if this technology is wildly successful?' has spawned a thousand science fiction stories: Daniel Mallory Ortberg famously suggested TV series Black Mirror is a response to the question, 'What if phones, but too much?' Alternatively, we might indulge some pessimism: How could this technology fail horribly? or How could someone abuse this technology? (We'll talk in chapter 6 about the dangers of technology being used for intentional harm.)

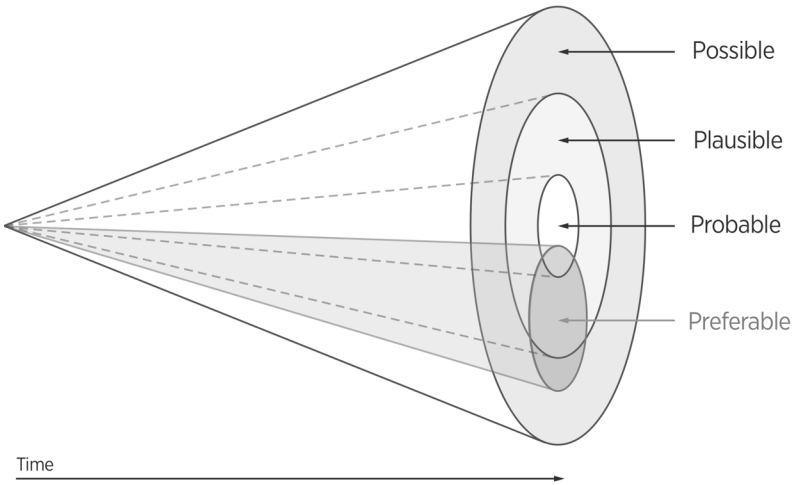
We can also use analogy to encourage moral imagination. Has this situation happened before, or in another field? What happened next? Could it happen here? Economic historian Carlota Perez argues that every technological revolution follows a tight script.[18] First, an 'irruption' phase, when a technology showing both promise and threat attracts heavy speculative investment. Then 'frenzy', a period of intense exploration, in which new markets explode into life and companies often cut ethical corners to make a quick buck. Eventually the bubble bursts and high-profile failures force regulators to step in. The momentum swings from finance to production as the technology becomes widespread; a phase of 'synergy'. Finally the revolution is complete and 'maturity' dominates. The market is saturated, awaiting the next disruption. The Gartner hype cycle traces a similar route for emerging technologies, from innovation through inflated expectation, through the trough of disillusionment, toward eventual stability.

To exercise moral imagination we can simply track our chosen technology along these likely trajectories, imagining what life might be like at each point. Today, for example, cryptocurrencies and machine learning are in the frenzy or inflated-expectations phase. It doesn't take a vivid moral imagination to picture what might happen when these inflated expectations burst.

Futuring

These set forecasts can be useful, but the world doesn't always follow neat blueprints. To stretch our moral imaginations further, we can learn from the field of futures studies. The central principle of so-called *futuring* is to see the future as plural. In the words of famed robot ethicist and futurist Sarah Connor, 'The future's not set. There's no fate but what we make for ourselves.'[19] The future isn't a mark on a map; it's the map itself. Collectively we get to decide which coordinates to head to.

A common model in futures studies is the *futures cone*,[20] which uses the analogy of light shone from a torch.



The *x*-axis represents time, so the torchlight represents potential futures. Note that the beam diverges from the present: next week is predictable; next century, rather less. Each cone of light represents a different level of likelihood, sometimes known as 'the 4 Ps'.

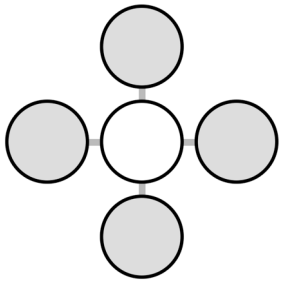
At the bright centre of the beam are the *probable futures*, the most likely projections based on what we know today. Gartner's hype cycle sits here, as does most day-to-day design work. Moving outwards, we come to *plausible futures*. These scenarios are less likely but still foreseeable: some corporations pay millions of dollars for insight into them. At the outer edges of the beam lie *possible futures*. Lying in

penumbra, these are harder to spot. Businesses typically aren't interested in these scenarios; instead, possible futures are the domain of what Anthony Dunne and Fiona Raby call 'speculative culture': science fiction, art, and games.

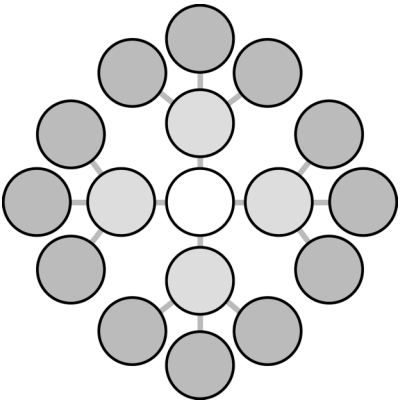
The final, and most important, cone depicts the *preferable future*. In a multitude of possible futures, some will be better than others. The preferable future is a value judgement; we have to consider the world we want and how we might get there. This ideal future might be highly probable, lying squarely in the middle of the beam, or an improbable wildcard found right on the edges.

Like any model, the futures cone has flaws. Design academic and critic Cameron Tonkinwise points out the direction of the beam depends on who's holding the torch: in an unequal world, everyone starts from a different 'now'. The idea of a preferable future is also loaded: preferable to whom? Nevertheless, the futures cone can be a useful prompt for strategy work and fostering moral imagination alike, illuminating various future trajectories and helping teams pick a preferred future to work towards.

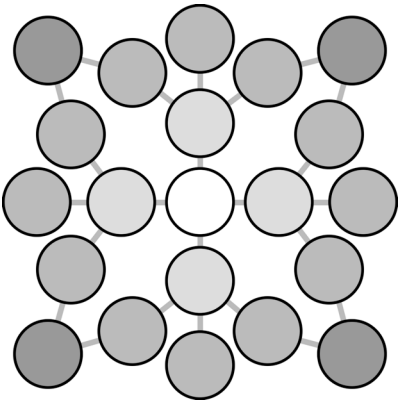
Another tool for teasing out potential futures and unintended consequences is Jerome C Glenn's *futures wheel*. In Glenn's words, the futures wheel offers 'a kind of structured brainstorming' about the future. We start with a root trend, such as our chosen technology; then, in a ring around the trend, we write some of its potential consequences.



This first step tends to draw out the probable futures, which are often interesting but rarely novel. So we now go deeper, picturing some potential second-order consequences of each new scenario, and adding these in a second ring.



Finally, we pair together interesting nodes from anywhere in the diagram, and imagine what might happen if both futures come true, recording this in a third ring. As the horizon expands, the chains of causality sometimes stretch, the possibilities becoming more imaginative, unexpected, and even apocalyptic. This isn't always the case, however: a third-order node can still represent a probable future if its predecessors are highly likely.



Running a futures wheel exercise with a tech team is usually as fun as it is eye-opening. It spawns some compelling stories about the possible impacts of our decisions, which makes it rich with ethical potential. Technologists don't often get a chance to think in this sort of speculative, even whimsical, way: the industry's focus on delivery tends to deter wild fantasies of the future. But futuring isn't about

accurate prediction so much as opening a team's eyes to possibility. Conjuring up shared visions of the future is the cornerstone of moral imagination, and a crucial way to expose unintended consequences.

> So why try to predict the future at all if it's so difficult, so nearly impossible? Because making predictions is one way to give warning when we see ourselves drifting in dangerous directions. Because prediction is a useful way of pointing out safer, wiser courses. Because, most of all, our tomorrow is the child of our today. Through thought and deed, we exert a great deal of influence over this child, even though we can't control it absolutely. Best to think about it, though. Best to try to shape it into something good. —Octavia Butler[21]

Design as provocation

Abstract, theoretical futures are lifeless and hard to picture; we need to experience them too. Moral imagination should involve emotion, not just logic. To awaken people to the potential consequences of their choices, we need to paint a vivid picture.

> It is not enough for a virtuous person to *intellectually* grasp her moral duty to extend compassion, or even to understand that it would be irrational not to do so. We must also find ways to *feel* compassion, which is an experience that goes beyond the intellect. —Shannon Vallor[22]

How can we bring these theoretical futures to life? How can we portray convincing scenarios that spark reaction, emotion, and debate? With design, of course. This is the premise of *speculative design*, pioneered by Dunne & Raby at London's Royal College of Art. This emerging branch of design focuses on how things *could* be, asking what-if questions to spark conversations and decisions about the futures we want.

Speculative design builds neatly on our futuring approaches. We can tease out potential futures and then illustrate a snippet of life in those futures. These *design fictions* can take the forms of short films, stories, games, comic strips, role-play, or objects: anything that builds a believable world.

A design fiction often features a hypothetical artefact, some believable prototype of the technology in question. I call this (with apologies) a *provocatype*. A provocatype isn't 'good' design: it isn't a thorough response to the brief, nor does it address every user need. Instead, it's designed to provoke conversation among stakeholders and potentially users too, if introduced with caution, NDAs, and the appropriate caveats. A provocatype creates a curious wormhole between design and research: it's a designed product that nevertheless exists mostly as a research probe. The big difference from regular prototyping is we make provocatypes with not just a problem-solving mindset but also a problem-creating mindset. If we're successful, a provocatype will spark better reactions than a hypothetical discussion would.

Let's see a provocatype in action, created by design firm The Incredible Machine for two Dutch energy clients. Their chosen future featured energy scarcity and electric vehicles sharing public charging stations. These are reasonable extrapolations from today: we could confidently call this a probable future. The designers chose to design a high-fidelity provocatype of the charging point itself.

The Incredible Machine, 'Transparent Charging Station'. Reprinted by kind permission.

As the purported user, you plug your charging cable (provided with the provocatype) into a free socket, tap your ID card to authenticate, and request your energy using the dials. The dotted display shows the charging queue and estimated completion times. But how

the cars get charged isn't the most interesting question. The provocatype's main role is to explore how an algorithm might prioritise energy when demand exceeds supply. It gives us an insight into an algorithmically driven future – we might call it an *algocracy* – a decade from now. This all hinges around the ID card, which the designers also prototyped.

Reprinted by kind permission.

Each user is given a card that relates to their position in society. A doctor's card lets them jump the charging queue, but carries a penalty for misuse. A recently released offender gets a probation ID, which gives them low charging priority and caps their energy use.

The designers, Marcel Schouwenaar and Harm van Beek, aren't proposing this as the optimal solution; instead, they've created an object that gets the right conversations happening and stimulates our moral imagination. We can't help but imagine what life would be like when our social status is wrapped into a digital ID. We see the potential positives – emergency services, for example, won't be unduly hampered by energy scarcity – but we also see how algocracy and the Internet of Things could reinforce social stratification and inequality.

Utopias and dystopias

Genevieve Bell points out a common pitfall with technological predictions: they often gravitate towards the extremes of utopia and dystopia.[23] We should be cautious of both.

Corporate vision videos usually depict a canonical capitalist utopia: gestural interfaces in gleaming offices, tediously perfect global collaboration. Ethically, these design fictions are empty: their provocatypes do very little provoking. But utopias can be dangerous, not just boring. The pursuit of a perfect society has at times been a gateway to extremism; the control required to make things just right can easily mutate into totalitarianism.

Dystopias are also seductive. Many designers will know the 'flip it' design game, in which participants imagine the worst possible solution to the brief, the idea being to then invert these ideas to uncover the principles of a successful design. Everyone has great fun drawing skulls and crossbones on things, and people often leave with surprisingly profound insights. Dystopias can indeed be powerful cautionary tales – Aesop's fables rarely had a happy ending – but dystopias can also be cynical and distant. They push away potential collaborators as much as engage them, and earn the ethically minded technologist a reputation as an obstructive fantasist.

The future usually follows a more nuanced path. When we're encouraging people to exercise moral imagination, we should steer clear of extremes. The ideal design fiction has a touch of moral ambiguity, hinting at good and bad alike. Speculative design intends to provoke responses, but it lets viewers construct those responses themselves, rather than forcing them to react in prearranged ways.

User dissent and crisis

Amid all this future-gazing we shouldn't lose sight of the human. Moral imagination should revolve around the people who'll live in our hypothetical futures and use our proposed technologies.

In *Design for Real Life*,[24] Eric Meyer and Sara Wachter-Boettcher recommend appointing a *designated dissenter*. This is a role of constructive antagonism, particularly useful in critique sessions. The dissenter's job is to challenge the team's assumptions, subvert deci-

sions, and lob in the occasional grenade of defiance. They might role-play as a user who refuses to provide the demanded data, or one who's insulted by the tone of an error message. Meyer and Wachter-Boettcher stress, however, that the role is best rotated: teams have a knack of tuning out a repeated naysayer, and too long wearing the robes of dissent can sour even the most charitable soul.

Design for Real Life also highlights moments of user crisis. The smiling personas taped to tech office walls depict users as happy and productive, although always incredibly busy. But real people aren't wooden archetypes. Our users include those who are coping with a job loss or bereavement, whose relationship is breaking down, or who are struggling with physical or mental illness. These moments are loaded with ethical significance. How we treat people at their most vulnerable is our deepest moral test, and as our technologies reach yet more of the planet, we'll have to support more people through these periods of crisis. The designated dissenter can help us imagine how these crises might occur in our chosen future, but there's also room for careful research, such as interviewing people who've experienced similar adversity. This research has its own delicate ethical issues, and is best left to trained researchers, perhaps supported by qualified counsellors for the most sensitive cases.

Redefining the stakeholder

Futuring and speculative design can reveal unintended consequences, but what about the externalities, the effects on people we've overlooked? As discussed earlier, economists tend to argue externalities need regulation, but the tech industry can and should try to reduce externalities too, by catering to a wider range of stakeholders.

Every business textbook offers a step-by-step guide to stakeholder analysis, but most only cover teammates or suspiciously homogenous groups like 'users' or 'residents'. This perspective, reinforced by the individualist focus of user-centred design, means we often overlook important groups. Stakeholders aren't just the people who can affect a project; they're also the people the project might affect. To force ourselves to consider the right people, try using a prompt list (see appendix) to capture a wider range of potential stakeholders, and use this as an input to futuring exercises and the design process.

Not all stakeholders will be welcome. In some cases, it might be worth including, say, a criminal, terrorist, or troll as a negative stakeholder – a *persona non grata*[25] – so the team can discuss how to actively reduce the harm this person can do. He may even deserve full persona treatment, with a name, an abusive scenario, and listed motivations to increase his profile within the team.

Stakeholders could even include social concepts: things we value in society but rarely consider within our influence, such as democracy, justice, or freedom of the press. As we now know, technology has the power to damage these ideas; explicitly listing them as stakeholders, or at least acknowledging their potential vulnerability, might help us protect them.

A Hippocratic Oath?

Moral imagination, futuring, provocatypes, and designated dissenters are all far from business as usual. Isn't there an easier way? Couldn't we start by creating a Hippocratic Oath for technology? This is an understandable and common question from people new to the tech ethics field; a written pledge seems like an obvious starting point, taking a cue from other disciplines.

A code of ethics might be useful at the right time, but this isn't a clear-cut ethical fix. The simplest argument is that it's all been done before. Dozens of previous attempts haven't bedded in; why would another be different? Designers will already know, for example, the First Things First manifesto of 1964, which argued designers should use their skills for moral good, not just for commerce. To be uncharitable, the manifesto's reprise in 2000 suggested the first incarnation had little long-term effect. In the domain of emerging technology, several codification efforts are already underway. IEEE's Ethically Aligned Design initiative has involved hundreds of experts, and details several key principles like human rights and accountability. Similar efforts include the Asilomar AI principles and the Barcelona Declaration. Professional organisations like the Association for Computing Machinery (ACM) also publish a code of conduct they expect from members.

These efforts, usually spawned from heavy consultation processes or elite conferences, can be bulky but are preferable to codes written

by a single author. The tech industry has seen a recent wave of what I call 'codes of reckons', simple bullet-point moral diktats from eminent technologists. These don't much help the cause of ethical technology. These codes' authors – unwittingly or otherwise – appoint themselves as ethical arbiters, projecting an unmerited stone-tablet authority. These documents lack public input and fall straight into the technocracy trap.

Ethical conventions don't themselves solve ethical problems: thorny moral questions still pervade medicine and engineering, despite the fields' prominent codes of ethics. Codes can offer some structure to ethical debate, but are usually too vague to resolve it. Consider two well-known maxims: the bioethics pledge 'First, do no harm' and Google's famous 'Don't be evil'. Although pithy, both statements are awkwardly imprecise in practice. What is harm? What is evil? Who decides? How do you resolve competing claims? To answer these questions we need more than a catchphrase; we need ways to properly evaluate ethical arguments. (We'll come to these in the next chapter.) Without this sort of moral framework, companies can choose any definition of evil or harm that excuses their chosen path. On its own, 'Don't be evil' means little more than 'Hey, ethics matters'. But we shouldn't be unfair. Acknowledging that ethics matters was itself something of a breakthrough at the time, and, while hardly a throwaway line, 'Don't be evil' was a small part of a Google staff manifesto rather than a formal corporate motto. In recent years it has become little more than a gotcha used by critics complaining about Google's mistakes. It's since been moved to the coda of a more detailed and more ethically useful code of conduct.

Codes of ethical technology also face enforcement problems. Professional bodies in many other disciplines have the power to disbar workers for malpractice, but since most technologists have no formal accreditation and membership of professional organisations is voluntary, codes of ethics in the tech industry are largely toothless.

Finally, codes are better at censuring bad behaviour than inspiring good. At worst, they can instil a checklist mindset, in which practitioners believe they can simply follow the numbered steps to pass ethical muster. Checklists have value but can be counterproductive if people fail to grasp the underlying spirit. In the field of web accessibility, the Web Content Accessibility Guidelines have been both a

help and a hindrance. They provide clear advice on accessible development, but have also caused some teams to misrepresent accessibility as a downstream box-ticking exercise: check your contrast ratios, tweak a few font sizes, and you're fully compliant. Ethical technologists know better. They know accessibility is really about who we deem worthy of our efforts; a commitment to treat every person as a person. We shouldn't mistake an ethical code for ethics itself. The conversation and the outcomes are what matters, not the paperwork. Ethics must become a custom, a way of thinking, a set of values held by all in the industry: as Cameron Tonkinwise calls it, *ethics as ethos*.[26]

Ethical infrastructure and diversity

It can be easier and more productive to codify ethics within individual companies, particularly if we can piggyback on existing policies. *Core values* – essentially a list of the company's stances and commitments – are a widespread and important vehicle for ethics, and usually carry senior support. Project teams can also create more localised rules, such as *design principles* that govern the design decisions within a particular product line or project. Strong core values and design principles taken to heart are powerful tie-breakers for ethical dilemmas: in the event of moral emergency, consult the agreed tenets for guidance. This means core values and design principles need to be specific. Some companies choose single-word values: Adobe's are 'genuine', 'exceptional', 'innovative', and 'involved'. Reflecting on the traits and qualities of a moral life can be important – it's the cornerstone of a branch of ethics we'll discuss later – but single-word values are just too slippery for a whole company. They leave too much unspoken, meaning people can twist them for their own purposes in a debate: 'How can you object to this tracking software? It's *innovative*!'

Sentences are better. Twitter's 'Defend and respect the user's voice' is a sound principle, although morally ambiguous: does this include defending hate speech? Ben & Jerry's core values are highly specific and even political: 'We seek and support nonviolent ways to achieve peace and justice. We believe government resources are more productively used in meeting human needs than in building and maintaining weapons systems.' This may be *too* specific – it's hard to

truly live by core values unless you can remember them – but it leaves no doubt about the type of company Ben & Jerry's wants to be.

According to researcher Jared Spool,[27] a good design principle is reversible. If you can flip the meaning and end up with a valid principle for a different team or time, you're being specific. 'Make it easy for users' is a platitude, not a design principle; the opposite would be absurd. The reversibility test doesn't fit core values quite so well. Sometimes it's helpful to explicitly support something that should be morally obvious – 'We care about the planet', for instance – but if in doubt, be specific.

Core values and design principles bolster a company's *ethical infrastructure*, as does team diversity. Homogenous teams tend to focus on the potential upsides of their work for people like them, and are blind to the problems they could inflict on a wider audience. The same divisions that pervade today's world are seen and even amplified in today's tech industry.

> If you live near a Whole Foods, if no one in your family serves in the military, if you're paid by the year, not the hour, if most people you know finished college, if no one you know uses meth, if you married once and remain married, if you're not one of 65 million Americans with a criminal record — if any or all of these things describe you, then accept the possibility that actually, you may not know what's going on and you may be part of the problem. —Anand Giridharadas[28]

Diversity and inclusion professionals often describe two dimensions to diversity: *inherent diversity* and *acquired diversity*. Inherent diversity refers to a group's innate traits, such as sex, orientation, and ethnic background, while acquired diversity refers to perspectives people have earned through experience. Both types of diversity can act as an early warning system for ethics. A team with broad inherent diversity will offer different perspectives and values, while people who are open to new experiences through, say, travel, literature, or languages generally find it easier to exercise moral imagination. While we should recognise the role of privilege – not everyone is lucky enough to see all the wonders of the world – actively absorbing new experiences typically strengthens one's ethical faculties.

> Fortunately, our powers of imagination can be increased. Seeking out news, books, films and other sources of stories about the human condition can help us to better envision the lives of others, even those in very different circumstances from our own. —Shannon Vallor[29]

Simply befriending and learning from people unlike ourselves also helps, building our mutual understanding and hence a sort of second-hand acquired diversity. The quest for diversity suggests we should also embrace interdisciplinarity. Slowly, the tech industry is learning that people from non-technical backgrounds, such as politics, law, philosophy, art, and anthropology, can bring huge value, not just in terms of different professional perspectives, but in much-needed acquired diversity. Long may the trend continue.