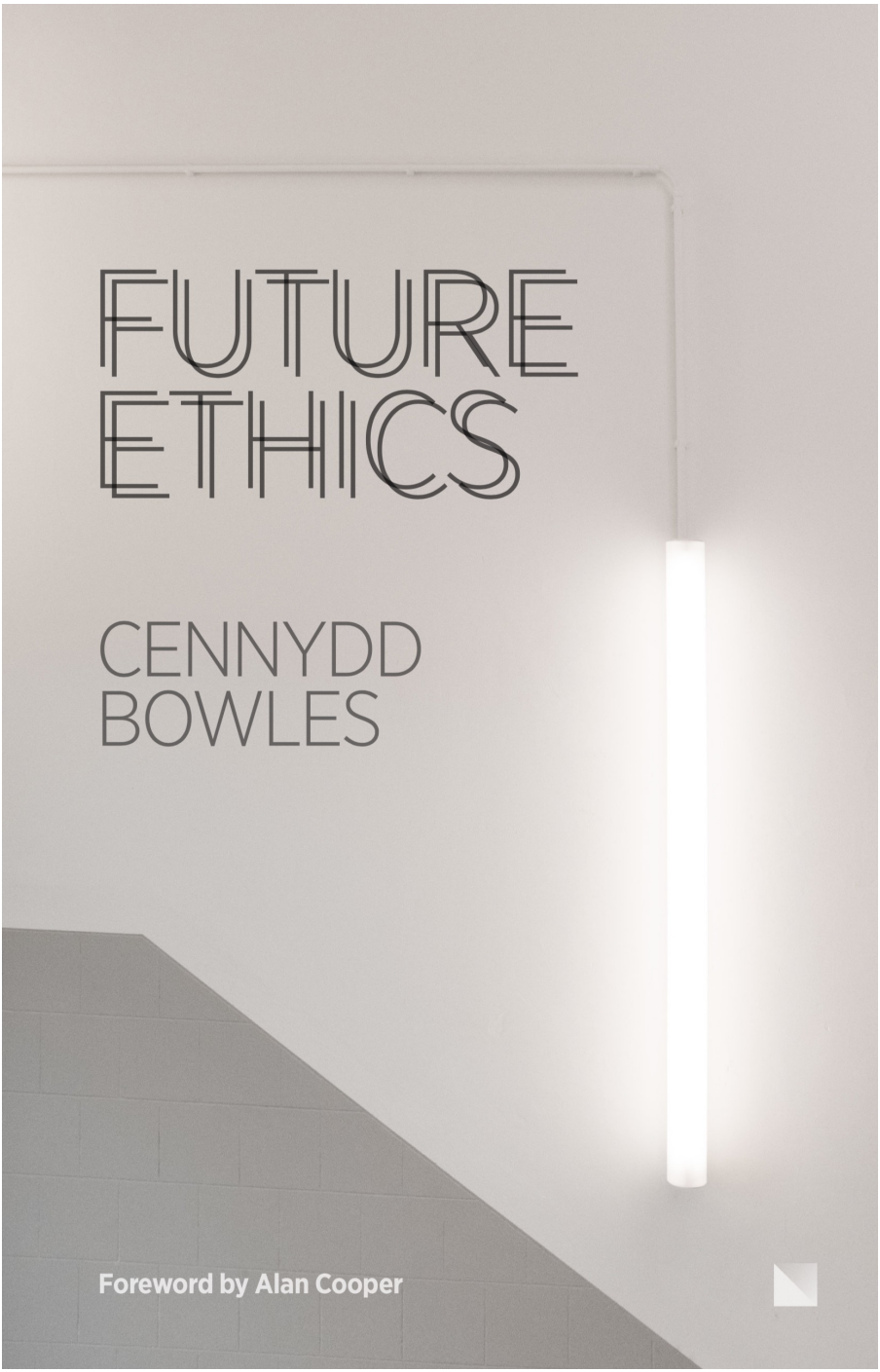


Free sample: Chapters 1, 2, and 3.

Future Ethics

Cennydd Bowles

NowNext
ETHICAL DESIGN & FUTURES

First published in 2018 by NowNext Ltd, The Old Casino, 28 Fourth Avenue, Hove, East Sussex BN3 2PJ, United Kingdom. nownext.press.

Cover image by Bernard Hermant. Typeset in Iowan Old Style.

ISBN 978-1-9996019-1-1

Calon lân yn llawn daioni,
Tecach yw na'r lili dlos:
Dim ond calon lân all ganu,
Canu'r dydd a chanu'r nos.

A pure heart full of goodness,
Is fairer than the pretty lily:
None but a pure heart can sing,
Sing in the day and sing in the night.

(*Calon Lân*, Welsh hymn)

CONTENTS

FOREWORD

Ethics has rightfully earned its reputation as a ridiculously boring topic. Much of the existing canon on ethics sounds to the modern ear like theologians debating the number of angels in heaven. It's either not relevant, or its prose is turgid and prolix, or both. Contemporary practitioners find little traction in the world of conventional ethical thinking. Yet there is much of value, if only it could be couched in relevant ways. And that is exactly what you hold in your hands. Cennydd Bowles has accomplished much in this single, readable, and relatable volume.

The dawning ubiquity of software and its data in every aspect of our world has opened a Pandora's Box of ethical questions. The questions aren't new – they have been debated for centuries – but they are assuming new shapes in their digital manifestations, and they are assuming new, far greater magnitude in contemporary social, economic, and political spheres.

Like Moore's Law, the world of software grows at an exponential pace, rather than linear. Growth at that speed means that signposts far off in the distance will be whipping past us much sooner than our intuition would suggest. The ethical questions raised by software's new capabilities are legion, and they demand answers now.

Armchair ethicists in tech circles discuss the trolley problem. This is a hypothetical exercise where one must decide – by switching tracks – whether a runaway trolley should kill a young

mother or an old man. It's true that makers of self-driving cars face a real world trolley problem, but average practitioners performing simple, day-to-day software design and development, are often blithely unaware of the ethical choices they make unconsciously. Yet these choices can have ramifications comparable to those faced by the inventors of poison gas or the atomic bomb. The frequency and consequence of ethical questions are far greater today than in the past.

In the industrial age, we allowed marketers to appeal to us with advertising design, product placement, and signage. The inherently sluggish atom-based world lulled us into believing this kind of persuasion to be harmless. Today, however, by analyzing your purchases and likes, Amazon can tailor its site for your personality and make you buy more stuff. While this is conceptually the same thing as cardboard signs in the grocer's aisle, its magnitude makes it a whole new world of morality. Some of the most fundamental questions about acceptable behavior are coming under scrutiny, and many of the answers need to change. And when the product being sold is a political party, or worse, a fear-based, totalitarian coup, change is imperative.

Software owned by someone else can gather data on you without you knowing it. Increasingly, it can do so without you even using it. Is that legal? Is that right? Who owns that data? What can they do with it? What rights do you have? What if that data is wrong? What if third parties make lasting decisions on that data? How do you even know?

Claiming that technology itself is morally neutral is a reassuring fig leaf worn by many technologists, but evidence against that idea has mounted to staggering proportions, and many digitally-savvy thinkers are confronting these apparently unprecedented moral dilemmas with fresh eyes.

Delegating authority to software without erecting robust feedback mechanisms is a common point of failure. Sure, the algorithm gives us an answer, but how do we know it's correct, and how do we fix it if not? For feedback loops to be effective, they have to be timely, and actionable, and some commitment has to be made to act on them. In pre-digital tech, the strong damping force of human participation in feedback loops helped to steady the process. Digital feedback loops

may be too efficient, causing the system to oscillate and fail. Mostly, though, feedback mechanisms are non-existent.

With today's tech tools, software often makes decisions on data gathered by some far distant app, owned by some unaffiliated company, at some far off time. Even conscientious organizations may find it impossible to verify that the choices their algorithms make are correct, let alone update them to do better in the future. The victim, typically just a person asking for a loan, or applying for a job, or trying to get veterans help, simply falls through the net.

A backlash is brewing against the digital world, and we could benefit from the lessons of history. We need useful principles spelled out today more than ever. As our digital artifacts mature, they exhibit capabilities unimagined in arenas unpredicted. Products intended to sell books have morphed into the arbiter of our built environment. Products intended for young adults to meet each other have morphed into brain-washing tools and tribal drums driving us to war.

Worse than a backlash is its opposite. Companies like Amazon, Facebook, and Cambridge Analytica are enthusiastically applying digital tools on everyone. These organizations are willing to use your information to make a profit without any consideration of your interests. And news outlets, whose industrial age distribution channels allowed some measure of detachment, are now virtual slaves to the click-through. Even mainstream news outlets like *The New York Times*, *The Atlantic*, and CNN have to exploit sensational 'news' in order to survive.

Governments, too, are playing this risky game. China is experimenting with a social rating system, like a credit rating, except it rates whether or not you are a good person. The potential for abuse is clear. The next steps are police robots that autonomously peer into your bedroom and decide if you are behaving properly.

At the start of this book, Cennydd gives us a survey of the ethical landscape. He articulates these important ethical principles in the context of modern tech, so they make sense to the contemporary practitioner. He then draws out extended examples of the growing ethical binds in the real world, and shows us how to apply the framework. What's more, he makes it interesting and applicable.

Because this is not a deterministic problem, there are no black-and-white answers. But there are many useful techniques for the

designer to master. The reader will be armed with a sense of mission, useful conceptual tools, and a map of the road forward.

We need good books about ethics now more than ever. Practitioners need guidance on how to think ethically, how to detect ethical choices, and how to resolve ethical dilemmas. That's exactly what this book is, and it's destined to become a well-thumbed classic.

—Alan Cooper,
21 August 2018,
Petaluma, California.

ACKNOWLEDGEMENTS

I'm indebted to:

My editor Owen Gregory, and my tech reviewers Thomas Wendt, Lydia Nicholas, and Damien Williams. No weak argument undented; no sweeping statement unchallenged; no scare quote unscrubbed.

Livia Labate, Eli Schiff, Paul Robert Lloyd, and Tom Hume for their invaluable reader feedback. Lou Rosenfeld, Richard Rutter, Abby Covert, Nick Disabato, and Brad Frost for publishing advice. Marcel Shouwenaar, Jeff Veen, Timo Arnall, Andy Cotgreave, and the Future of Life Institute for photo permissions. Christina Wodtke, Azeem Azhar, and Jon Kolko for the flattering blurbs, and Alan Cooper for his gracious foreword.

Attendees of the Juvet AI retreat for the inspirational conversations and childlike Borealis wonder, and everyone else working in the field of emerging technology ethics. Your work is vital and deep; I hope I've done credit to your ideas.

My family and friends. Thank you in particular to Sascha Auerbach and Andrew Fox for their ongoing (liquid) support.

Finally, my wife, Anna. My guiding star.

CHAPTER 1

TROUBLE IN PARADISE

The utopian dreams of early cyberspace didn't come true. Eden was rezoned, walled off. The lemonade stands grew into colossal malls; disinformation and deceit polluted the global agora. After fulfilling their promise to demolish old hierarchies, technologists erected new towers and fiefdoms in their place.

Industry orthodoxy – the 'Californian Ideology' described in Richard Barbrook and Andy Cameron's influential essay – sees technology as the solution to any problem. To Silicon Valley's cheerleaders, technology is intrinsically empowering, so laden with good that harm is almost unthinkable. A spirit of exceptionalism courses through the community's veins: believers see themselves as beta testers of a brave new world, and regard existing social structures, norms, and laws as anachronisms, inconveniences best routed around. Technologists have learned to build first and ask questions later. Lean startup, tech's predominant ideology today, is vehemently empirical. It argues that we're so swept up in change it's futile to predict the future; instead, we should prioritise validation over research and learn through making. Build, measure, learn, repeat.

This approach has brought bold innovation to stagnant fields, but when technology becomes an answer to any problem, it should be no surprise that 'Can we?' overtakes 'Should we?' Just as promised, technologists have moved fast and broken many things. The industry's repeated missteps – racist algorithms, casual privacy abuses, blind

eyes turned to harassment and hate – have eroded public faith and prompted the media to label technology a danger as often as a saviour. Tech employees may be surprised to find themselves in the crosshairs. Most genuinely want to improve the human condition, or at least tackle interesting problems, and have good intentions. The industry's problems are mostly down to negligence, not malice.

An ethical awakening is long overdue. Technologists are rightly starting to question their influence on a world spiralling off its expected course, and as the industry matures, it's natural to pay attention to deeper questions of impact and justice. As sociologist Richard Sennett points out, 'It is at the level of mastery […] that ethical problems of craft appear.'[1]

This focus coincides with growing public disquiet and appetite for ethical change. Consumers want to support companies that espouse clear values: 87% of consumers would purchase a product because a company advocated for an issue close to their hearts.[2] Emerging technology raises the stakes further. Over the coming decades, our industry will ask people to trust us with their data, their vehicles, and even their families' safety. Dystopian science fiction has already taught people to be sceptical of these requests; unless we tackle the ethical issues that are blighting the field, this trust will be hard to earn.

Instrumentalism, determinism, and mediation

As our first ethical step, we should abandon the comforting idea that technology is neutral. This *instrumentalist* stance argues technology is just a tool, one that people can use for good or misuse for harm. Instrumentalists argue that since bad actors will always twist technology for evil, the only ethical recourse is to educate and plead for proper use. This deflects responsibility onto the user, allowing technologists to wriggle off the moral hook. We all know one popular instrumentalist refrain: 'Guns don't kill people; people kill people.'[3]

The opposing view – *technological determinism* – argues that technology is anything but neutral; instead, it's so powerful that it moulds society and culture, acting more as our master than our servant. Determinism pervades both science fiction and academia, and has even begun to seduce the media; gleeful reports on technology's

brewing dominance over mankind litter today's front pages. Politicians are starting to catch the determinist bug too, declaring that technology will define the twenty-first century.

Instrumentalism is handy for shutting down critique: if technology is just an inert tool, it has no social, political, or moral effects. However, the industry has been obtuse in clinging to this view; tech marketing suggests the industry is well aware of its potential impacts. Technologists often describe their lofty goals with deterministic language – Democratise! Transform! Disrupt! – but fall back on instrumentalist defences to ethical issues: we truly regret this disturbing case, but we can't be held liable for misuse. In other words, technology will change the world, but if the world changes, don't blame us.

Technology's harmful impacts make instrumentalism unsustainable; even the supposedly benign search engine has reinforced bias and devalued trusted information sources. Opposing the neutrality myth is hardly a new stance. In 1985, tech historian Melvin Kranzberg presented six laws of technology. Law one: 'Technology is neither good nor bad; nor is it neutral.' But in rejecting instrumentalism we shouldn't necessarily leap to determinism. Putting technologists at the centre of the universe isn't healthy for an industry in dire need of humility, and determinism can curiously downplay technologists' ethical responsibilities. If we see technology as an unstoppable social force, we might conclude it's outside our control.

Tech philosopher Peter-Paul Verbeek suggests a third perspective – *mediation theory* – that neatly melds the competing views of instrumentalism and determinism.[4] For Verbeek, technology is a medium through which we perceive and manipulate our world. Glasses help us see and understand our environments; hammers help us build shelters and sculptures; cameras help us recollect and share our memories. Perhaps it's futile to separate technology from society. We don't fully control tech, nor does it fully control us; instead, humans and technologies co-create the world. An anecdote from Kranzberg about the violinist Fritz Kreisler shows this combination at work:

> A woman came up to [Kreisler] after a concert and gushed, 'Oh, Maestro, your violin makes such beautiful music.' Kreisler picked up his violin (a Stradivarius, no less), held it to his ear, and said, 'I don't hear any music coming out of it.' You see, the beautiful music coming out of the violin did not come from the instrument, the hardware, alone; it depended upon the human element, the software.[5]

Only the violinist – a hybrid of the violin and the human – could create such memorable music (although we can blame Kreisler alone for the arrogant witticism).

Barriers to ethics

What does this mean for ethics? If humans and technology act in tandem, we can't claim technology is ethically inert, but neither can we separate it from human action. The ethics of technology becomes the ethics of everyday life. But as a conversation topic, ethics doesn't always spark enthusiasm. All those pointless thought experiments and dusty Greeks! Not to mention the definitional pedantry: what's the difference between ethics and morals, anyway? Perhaps your mind will wander back to high-school religious studies or civics lessons: isn't ethics about society's expectations and morality something more personal and innate? There's a deep philosophical rabbit hole here, but happily we can choose to sidestep it. Most (but not all) modern philosophers see no big difference between morals and ethics, and use the terms interchangeably. I will too.

Whatever the label, ethics matters more outside the classroom. Ethics is a vital and real topic, nothing less than a pledge to take our choices and even our lives seriously. This commitment is especially important for designers. Design *is* applied ethics. Sometimes this connection is obvious: if you design razor wire, you're saying that anyone who tries to contravene someone else's right to private property should be injured. But whatever the medium or material, every act of design is a statement about the future. Design changes how we see the world and how we can act within it; design turns beliefs about how we should live into objects and environments people will use and inhabit. In choosing the future they want, designers discard dozens of alternative realities, which pop briefly into existence through proto-

types or sketches, but perish in the recycling bin. As one memorable quote proclaims, 'Ethics is the aesthetics of the future.'[6]

Bringing up ethics in the workplace often prompts two objections. First, some people claim ethics doesn't belong in industry, and that acceptable behaviour is for the market or the law to decide. This is a political idea, and its weaknesses should be clear to anyone who disputes its libertarian premise. A market that ethically self-corrects requires perfect information and full agreement on what's right and wrong. Customers can only punish ethical overreach if they know and understand what companies are doing, and if they agree it's unethical. But technology acts invisibly, often with dubious consent, and typically using a dialect only a few can speak. The general public has no idea what sorts of unwelcome acts are happening inside their gadgets. The idea of market self-correction is a fantasy.

The claim that the law is the best ethical arbiter is particularly wretched; it essentially argues we should allow all behaviour except the criminal. Ethics should be about living our best lives, not seeing how low we can sink. And laws themselves can be morally wrong; sometimes brave people have to disobey unjust legislation to spark ethical change: just ask Rosa Parks. Even if we ignore these arguments, for law to be an appropriate substitute for ethics in tech, we'd have to find legislators who deeply understand technology. History tells us these individuals are sadly rare.

The second common objection to business ethics is that it will hamper innovation. Sometimes that's true. Pausing to take moral stock will indeed extinguish some potentially harmful ideas, but an enlightened company should be grateful for the intervention. Ethics isn't just a drag on innovation; properly handled, it can fertilise new ideas as well as weed out bad ones.

This book

While it's heartening that technologists are finally taking ethics seriously, we shouldn't believe we're the first on these shores. Sadly, the philosophers, academics, writers, and artists who have studied the topic for decades aren't yet taken seriously within industry; tech culture prizes intelligence but is doggedly anti-intellectual. In turn, academics complain about practitioners' hubristic ability to run

repeatedly into the same old walls, while being paid handsomely to do so.

As a working designer, not an ethicist, I'm writing this book for my peers in the tech industry. While the book owes deep gratitude to those who have paved the way, and won't shy away from complex ideas, I'll try to always translate theory into application. That said, a manual for ethics is an oxymoron; if you're after bullet-point instruction, you'll be disappointed. No one can answer ethical problems for us; we have to think them through for ourselves, and there's usually more than one answer. To paraphrase Caroline Whitbeck[7], ethical issues are like design briefs: there are often dozens of viable solutions, each with their own trade-offs. This doesn't mean there are no wrong answers, however. Ethics is beset with pitfalls and fallacies; we'll highlight the most common ones as we go.

Some politics is inevitable in a book like this, since ethics and politics are naturally entwined. The breadth of human opinion is reflected in the complexity of ethics; people's moral views tend to inform their political views, and vice versa. Those on the left might favour moral stances that prioritise social good, while those on the right may prefer perspectives that support individual sovereignty and autonomy. Personal experience also bears a strong imprint: a victim of robbery will probably feel more strongly about theft in future, whether consciously or not. It would be disingenuous for me to disguise my personal and political leanings in the name of false objectivity, but I'll try to avoid cheap point-scoring and instead give you tools to work through ethical arguments for yourself. You may even find that thinking deeply about ethics influences your views on broader society.

Thank you for taking an interest in forging a better tech industry; I hope this book will give you both the theory and practical advice you need to do just that. Let's get started.

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.)

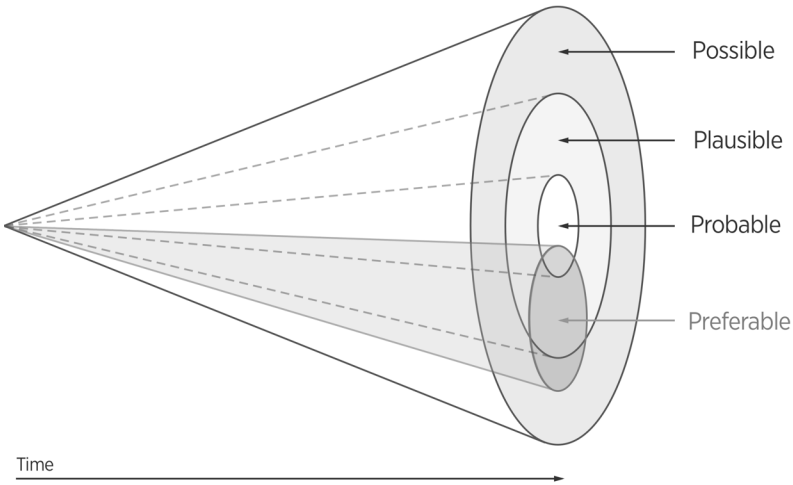
We can also use analogy to encourage moral imagination. Has this situation happened before, or in another field? What happened next? Could it happen here? Economic historian Carlota Perez argues that every technological revolution follows a tight script.[18] First, an 'irruption' phase, when a technology 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'.

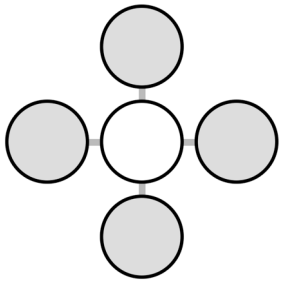
At the bright centre of the beam are the *probable futures*, the most likely projections based on what we know today. Gartner's hype cycle sits here, as does most day-to-day design work. Moving outwards, we come to *plausible futures*. These scenarios are less likely but still foreseeable: some corporations 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.

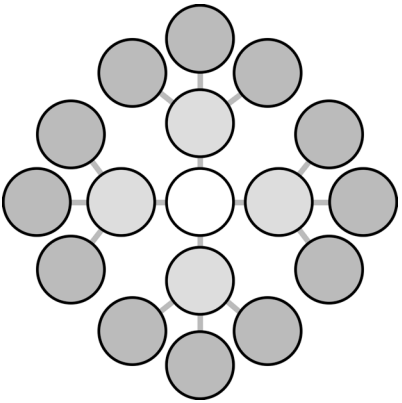
The final, and most important, cone depicts the *preferable future*. In a multitude of possible futures, some will be better than others. The preferable future is a value judgement; we have to consider the world we want and how we might get there. This ideal future might be highly probable, lying squarely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

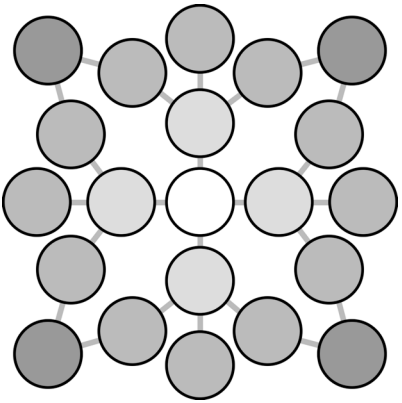
Another tool for teasing out potential futures and unintended consequences is Jerome C Glenn's *futures wheel*. In Glenn's words, the futures wheel offers 'a kind of structured brainstorming' about the future. We start with a root trend, such as our chosen technology; then, in a ring around the trend,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

CHAPTER 3

PERSUASIVE MECHANISMS

In his influential essay 'Do Artifacts Have Politics?',[1] Langdon Winner concludes that yes, they do. He challenges the instrumentalist idea that objects are just inert products of the social forces that created them, and argues instead for a wider view: that objects themselves affect how power and authority are distributed, and how societies behave.

The essay examines, among other things, the overpasses that span Long Island freeways. Winner claims New York planner Robert Moses built these bridges unusually low to achieve 'a particular social effect', namely segregation. Poor residents, and non-white residents in particular, typically travelled by bus at the time; since these buses couldn't fit under the bridges, these people were effectively excluded from Long Island's beaches. Although Winner's account is now somewhat disputed, it still shows that even hulking masses of concrete and steel can enforce social change.

Designers are already well aware of the power of objects. For decades, graphic designers have tried to change public attitudes and behaviours, devising not only the appealing cigarette packet but also the calming hospital signage for the smoker's final days. Technological objects – computers, handsets, gadgets – can be particularly potent at kindling new desires and moulding behaviours: nothing so full of language, light, and energy could ever be inert.

Coercion vs. nudging

Designers sometimes embed moral decisions into the environment by force, meaning the user has to comply with the designer's wishes. Speed bumps force drivers to brake; safety catches make it harder to accidentally fire a weapon. Digital technology often similarly constrains user choice. We often hear that design is a conversation with the user; in tech, the conversation is woefully one-sided. In the words of former Google design ethicist Tristan Harris, 'whoever controls the menu controls the choices'. Short of learning a programming language, you can't make a computer do anything its interface doesn't allow. Design decisions, therefore, give technologies the power to enforce behaviour – and hence moral conduct – in the designer's absence.

Coercion might seem unethical since it limits people's free will, but plenty of our society relies on coercion and compliance, particularly our laws. Coercion instead affects where moral responsibility lies. You aren't responsible for behaviour that isn't freely chosen; we don't blame someone forced at gunpoint to commit a crime. Responsibility lies instead with the coercer: a designer must take responsibility for any decisions they force a user into. Military trials have made it clear, however, that orders from superiors don't count as coercion; soldiers can refuse to comply with unlawful or immoral orders, although they may pay a heavy price for it.

There are subtler persuasive arts than blunt coercion. *Nudge theory*, popularised by Richard Thaler and Cass Sunstein, tries to steer behaviour through simple changes to defaults and framing. Nudge has flourished in the public sector – opt-out policies for organ donations, electronic signs that smile or glower at a passing driver's speed – but Silicon Valley is also fond. Nudge doesn't trespass on individual freedoms the industry holds dear, but is still a potent technique for liberating people from their money or time.

Nudgers are quick to point out they merely massage available options, rather than reduce them. This paints nudge as a technique of persuasion, not coercion. However, persuasion still carries ethical concerns. As Daniel Berdichevsky and Erik Neuenschwander, early theorists of persuasive technology, note, 'Persuaders have always

stood on uneasy ethical ground. If a serpent persuades you to eat a fruit […] does culpability fall upon you or upon the serpent?'[2]

Even if persuasion isn't the explicit plan, design always influences behaviour. A design is successful if it steers the user to the right information or the next step in the process; enlarge a button to make it more visible and you'll find more people will push it. All target-driven design, therefore, is persuasive design. Any team with performance targets will try to manoeuvre user behaviour to reach company goals. This means we can't simply pledge to never practice persuasion: we'd have to quit design altogether. Instead, we have to dive in with intent, acknowledging our responsibilities and choosing how to address the ethical challenges.

Dark patterns, attention, and addiction

In a rational world, persuasion would be simple: outline the benefits and costs for each option and trust the user to make the right choice. No such luck. Persuaders must also appeal to bias and emotion, to what we may consider human weakness. Yet behavioural design often wears the cosy clothing of paternalism, faintly patronising but broadly beneficent. Nudge does exploit human weakness, but nudgers would argue they do it so we can overcome that weakness. Don't we all want to live healthier, more responsible lives? Persuasion itself is therefore positioned as a benign tool that elevates us all, helping people ascend the face of Mt. Maslow and reach an enlightened summit.

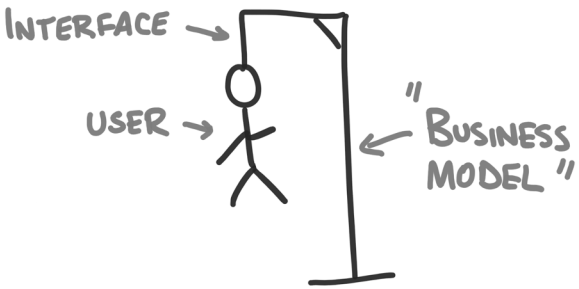
But persuasion can corrupt humanity as much as ennoble it: the techniques that can help people lose weight can also be used to encourage them to skip voting in the next election. In the tech world, unethical persuasion often takes the form of a *dark pattern*,[3] an intentionally deceptive interface that exploits cognitive weakness for profit. Most dark patterns today are extortive nuisances – fake scarcity on hotel sites, bait-and-switch subscriptions – but the dark pattern becomes more threatening as technologies become embedded in everyday life. Persuasive technology may fade from sight, but its force fields grow ever stronger.

Persuasive techniques can be pointed back at themselves: technology can persuade us to use technology more. Social media addiction has become a full-blown moral panic, fuelled by tabloid horror

stories and pop-sci literature that reduces deep research to dopamine anecdotes and frontal lobe neurobollocks. Social media joins a rich canon of moral scourges: books, newspapers, and gramophones were all, in various centuries, linked to the certain downfall of social order.

It's common to blame Silicon Valley business models for the addiction crisis. Free services aren't really free, we're told; instead, users cough up an alternative currency – attention – which is monetised through advertising. In sectors like free-to-play gaming, this time/money equivalence is painfully literal: watch an advert and earn fifty coins. But we can trace a thirst for attention back through advertising, media, and even religion: the attention economy isn't just a consequence of search engines and social networks.[4]

The modern adage 'If you're not paying for the product, you are the product being sold' is facile. It implies that paying with attention is less ethical than paying with cash, a discriminatory position that suggests the poor are somehow less deserving of technology than the rich. It also ignores the market's habit of demanding paid and free services alike grab as much attention as they can. In July 2017, a 17-year-old boy in Guangzhou suffered a minor stroke after a forty-hour stint playing Tencent's mobile game Honour of Kings. Tencent responded bravely, announcing they would limit children's playing time. The company's stock sagged by 5.1%. Even subscription services, brimming with recurring revenue, brag in earnings calls about daily active users and time-in-app metrics. Even if you *are* paying for the product, your attention is still being sold. A cynic could go further, arguing that ensnaring the user in a business relationship is the whole point of experience design; an assertion sardonically illustrated by Jeff Veen.



Redrawn and used with kind permission of Jeff Veen.

How harmful the attention economy is depends in part on whether its effects are zero-sum. Screens that only drag us away from other screens are relatively harmless, but if technology distracts us from what ethicists call 'the good life' – a vague concept, but one that might reasonably include family, friends, productive work, and self-improvement – technology becomes a force of alienation, as predicted decades ago by philosopher Karl Jaspers.

It now seems Jaspers' fears are being realised. Netflix CEO Reed Hastings has claimed 'we're competing with sleep';[5] the average American spends 3.1 hours on a mobile device each day, compared to just eighteen minutes in 2008, with no corresponding drop in desktop use.[6] We are all Sisyphus reborn, reducing unread counts each day until, in the words of Ian Bogost, 'conflict and exhaustion suffocate delight and utility'.[7] We are right to fear technology that erodes the rest of our lives.

Addiction concerns will grow as tech firms compete for long-term attention and the rich advertising seams dominated by TV and movies, and as more immersive technologies reach our homes. Virtual and augmented reality both offer the potential of an irresistible hyperreality, as foretold throughout science fiction. The common theme of these stories is, per Jaspers, alienation from authentic human experience: a science fiction character who gives up on the physical world rarely enjoys a happy ending. However, the threat is no longer purely fictional. The alarming phenomenon of *hikikomori* ('withdrawing inward') has seen hundreds of thousands of young Japanese men shrinking from society. Unlike the couch potato, the hikikomori does not embrace idleness for its own sake; rather,

unable to cope with the stresses of the external world, he retreats into himself. Many hikikomori find themselves drawn into immersive media or games. Although we mustn't confuse causation and correlation, it's clear immersive technologies, along with automation, the collapse of local retail, and ageing populations, could cause people to withdraw from society.

Experimentation

Tech companies have so embraced behaviour change that they trial countless designs to find the most persuasive variants. Perhaps a larger Buy Now button will increase sales, or a different voice assistant script will encourage more queries. This empirical approach is reinforced by technologists' love of the scientific method, instilled in their STEMish undergrad days, and the rise of lean startup. Lean enthusiasts contend iteration is the best route to product–market fit: experimenting with, and on, users is celebrated as a natural step in this process.

Any project that learns from user behaviour is a user research project, yet the industry has tacitly chosen to exempt experimentation from research ethics. Users are given no right to withdraw from studies. Children are routinely included in experimental populations. Informed consent is brushed aside, supposedly replaced by an excusatory sentence in the terms of service. An institutional review board (IRB) would rebuff academic research this sloppy, but the industry argues this level of ethical oversight would neuter innovation. While regulators turn a blind eye, companies experiment recklessly on users. Experimentation can be a powerful way to test product improvements, but in some companies it has mushroomed beyond interface tweaks into psychosocial research, with sometimes shocking disregard for public wellbeing: Facebook's 'emotional contagion' study, which rigged the News Feed of 689,000 users to learn whether it affected people's moods, is a notorious example.[8] The research ran under the auspices of Facebook's own data policy rather than that of an IRB. Facebook presumes a user accepts the policy as a condition of using the service, but most users never open it, let alone read it. It's laughable to claim the policy ensures informed consent. No opt-out was offered, and the researchers seemed to ignore the

study's potential impact on users with depression, despite the resulting paper mentioning the condition as a focus of prior research.

Many tech companies responded to the resulting outcry with a shrug, claiming this is simply how technology works. OkCupid boasted 'We Experiment On Human Beings! (So does everyone else.)', and the Facebook researchers expressed their shock at the reaction in keynote speeches. These blasé defences of experimentation are rooted in the dehumanising effects of scale and the industry's quantitative bias. Marry tech companies' enormous reach with a belief that progress must be measurable – that objectives and key results (OKRs), active users, or conversion rates are the only worthwhile barometers of success – and a culture of target-chasing often wins out. Gradually, users become not *raisons d'être* but subjects for experimentation, means for teams to achieve their own goals. We start to see not customers, but masses.

Persuasion and power

The dynamics of persuasion are often political. Even the ostensibly benign nudge has partisan effects. Citizens and politicians alike find the idea of nudging more ethical when the examples given align with the subjects' politics. Even libertarians, typically wary of nudge's overbearing tendencies, set aside their scepticism when they approve of the nudger's goal.[9]

At the height of the 2010–12 Arab Spring, technology felt emancipatory, a positive force for uprooting hierarchy and oppression. How naive that now seems. Today, the internet has become the key battleground for political persuasion, propaganda, and disinformation. By manipulating information channels like social media, parties and nations can jostle for narrative supremacy: a strategy sometimes known as *influence operations*.

The web's structure helps these efforts. Respected news sources, niche publications, and propaganda factories can all reach global audiences; all are a single HTTP request away. Modern conspiracies also look as legitimate as any respectable story. In years past, we could identify crank literature by its format – ugly scrawls and bad photocopying – but now templated publishers like Medium or Squarespace allow anyone to publish information in a credible format. Disinforma-

tion is aesthetically equivalent to a legitimate press release, and a social media post from an lone conspiracist looks the same as an official announcement from the BBC.

Hypertext represents knowledge in ways that encourage exploration, fragmentation, and reassociation. Hypertext therefore tends to break apart centralised, linear narratives and encourages instead *apophenia*, a habit of imposing relationships on unconnected things.[10] Conspiracies bloom in the dark, in the gaps between trusted information, allowing the powerless to explain away their lack of agency. In this battle of the timelines, no matter how outlandish the narrative, there are people ready to believe. Furries, flat-earthers, and fascists alike can stitch together their own narratives from the digital fragments of the web, and broadcast them into their communities.

From the 'paranoia tourism' of Pizzagate to Reddit's terrible unmasking of the wrong Boston bomber,[11] extremists and ideologues have smartly exploited the public using Silicon Valley's persuasive infrastructure. However, tech companies have denied any role in political persuasion, sticking to their instrumental excuse: we're neutral platforms, not media companies. This is a feeble defence. No industry that spends millions lobbying for deregulation can claim political neutrality. Facebook's denial of political influence particularly galls when their partnership puff pages boast of an 'audience-specific content strategy to significantly shift voter intent.'[12]

There are several reasons why tech companies have been slow to stamp out harmful and misleading information. Identifying this content is certainly difficult; no company will hire vast fact-checking teams, and automated efforts will throw up plenty of false positives. But social networks fundamentally didn't much care about the quality of information shared, so long as it was shared; almost anything that moved the needle was welcome. Companies only started paying due attention to the propaganda externality once politicians held them accountable and dragged executives in front of committees and tribunals.

In retrospect, the industry's failings on propaganda have familiar origins. In their rush to build, technologists didn't consider how the structures and affordances of their new systems might have unintended consequences. Tech teams failed to mitigate the risks,

meaning society has to bear them instead, in the form of resurgent extremism and conspiracy.

Automated persuasion

Automated persuasion – artificial agents with their own forms of algorithmic inducement – may pose an even larger menace to truth and democracy. We mustn't repeat the same mistakes.

Bots are already a viable persuasive threat. Analysing the digital ecosystems of the 2016 US election, Berit Anderson and Brett Horvath uncovered 'a weaponized AI propaganda machine'[13] that hinged on Cambridge Analytica profiling, automated scripts, and a deep network of propaganda sites. Relying on disenfranchised people's appetite for conspiracy, disinformation accounts on social media whipped up dissent in sympathetic communities. Oxford academics observed a similar, albeit smaller, pattern during the Brexit referendum: Twitter propaganda accounts previously used to skew opinions on the Israel–Palestine conflict were given a coat of British nationalist paint and flung into the new debate.[14]

Reporting of the Brexit and Trump campaigns has been inconsistent: many so-called bots were instead paid trolls, apparently part of rival states' influence operations. Sloppy labelling is understandable – people are still scrambling to understand how our technologies are being turned against us – but automated persuasion has undoubtedly played a part in recent political upheaval. It will only become more influential.

Persuasion is an ideal candidate for machine learning. We can define simple metrics we want to maximise (more followers, clicks, and retweets seem like decent proxies for influence), offer a wealth of behavioural data to mine, and propose hundreds of potential parameters to tweak. Political bots can trial dozens of conversational approaches, hashtags, and slogans; a design bot can test countless interface permutations to induce a potential customer to buy. Amazon already employs automated nudges at vast scale.

> Through our Selling Coach program, we generate a steady stream of automated machine-learned 'nudges' (more than 70 million in a typical week) – alerting sellers about opportunities to avoid going out-of-stock, add selection that's selling, and sharpen their prices to be more competitive. These nudges translate to billions in increased sales to sellers.[15]

Georgia Tech researchers found people were surprisingly susceptible to machine persuasion in an emergency.[16] Test participants were greeted by a crude robot and instructed to follow it to to the lab; half of the time the robot took wrong turns, to give the impression it wasn't exactly a competent navigator. Midway through the study, experimenters flooded the adjoining corridor with fake smoke, setting off a fire alarm. During the phoney emergency, not one participant escaped the way they'd come in, or rushed to an emergency exit: they all followed the robot's instructions to head for a back room, even if they'd seen the robot make mistakes or break down earlier. Participants probably knew the emergency was faked – an IRB would have ethical misgivings about a study that made people genuinely fear for their lives – but people still 'overtrusted' the machine far beyond the researchers' expectations.

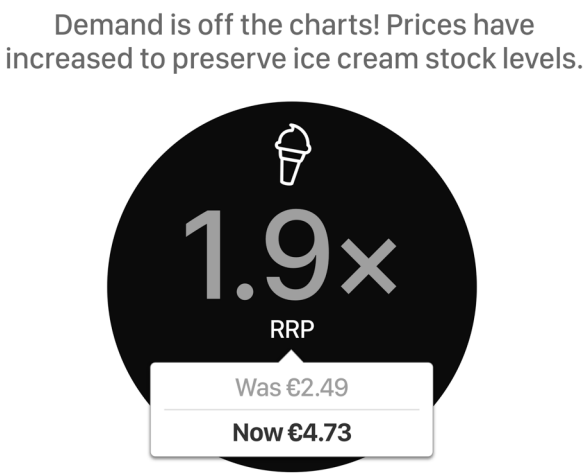
Emotional, or affective, technology will further sharpen the persuasive toolkit. London's new Piccadilly Lights billboard uses hidden cameras to deduce the gender, age, and mood of people in the vicinity, so it can serve ads appropriate to the audience: the public realm becomes a data mine. The endgame for this scenario – an artificial agent that can not only read gesture, intonation, and body language but mimic it in its responses, adapting not just what it says but how it says it – will be a formidable manipulator.

Automated persuasion is structurally quite different from existing forms of mass persuasion like advertising. Persuasive algorithms can respond to change rapidly, learn from millions of successes or failures elsewhere in the network, and can be highly personalised. Given enough data and training, an algorithm can present a compelling, tailored message to every individual; one-size-fits-all marketing gives way to a system that pushes only your most sensitive buttons. Legal scholar Karen Yeung argues automated persuasion is so unlike its monolithic predecessors that it deserves a new name: *hypernudge*.

Hypernudges could transform persuasion into duress. In a 2017 study, researchers created a 50% purchase uplift simply by tailoring Facebook ads to users' inferred personality types.[17] Clearly, this profiling will become more sophisticated and tied to more persuasive messages in future. If we can exploit someone's weaknesses on cue, when does a nudge become a shove?

Perhaps the biggest ethical challenge of hypernudging is its invisibility. There's no way to tell whether a camera is feeding a persuasion engine; soon we won't know whether a help desk is staffed by humans or hypernudging algorithms. A power imbalance is implicit in connected technology: the public has no insight into the network's machinations and no recourse against exploitation. Historically, mass persuasion was homogenous and visible: everyone saw the same newspapers, ads, and political broadcasts. This meant they could be critiqued. People could object en masse to misleading or unethical persuasion, and authorities could demand a message be withdrawn. But invisible hypernudges allow less scope for protest. With persuasive content tailored to the individual and delivered on a personal device, it will be harder to unify in opposition, and authorities will struggle to take corrective action.

Price also has persuasive effects. Dynamic pricing is hardly new – airlines have been at it for years – but it could soon hit a wider range of industries. Algorithms will be able to fine-tune prices to preserve stocks, manipulate demand, and, of course, extract maximum profit. Networked electronic price tags allow retailers to adjust prices instantaneously, which reduces waste and point-of-sale admin, but also allows retailers to spike prices at times of peak demand: surge pricing at the gelato counter.



In theory, algorithmic pricing could have some social benefit. Matching price to someone's ability to pay could help address inequality, and explaining price variations could illuminate opaque supply chains: your latte's more expensive this week thanks to storms hitting production in Vietnam. But customers typically see price discrimination as unfair, meaning today it often meets fierce backlash. Just ask Orbitz, who were caught serving Mac users higher hotel prices, on the presumption of higher income.[18]

If algorithmic pricing does become widespread despite this opposition, the public's only recourse might be to collectively confuse or skew price signals: *price hacking*. Price hacking has already been recorded among Uber drivers: by colluding to go offline in unison, drivers create a supply deficit that instigates surge pricing. In a world of mundane algorithmic pricing, customers may find themselves following suit, abstaining from a product to crash prices, then stockpiling en masse. This might in turn spark secondary resale markets, or even some sort of futures trading. Friends will pool pricing information and ask whoever gets the best price to buy on others' behalf. Algorithmic pricing may even create de facto cartels by accident: if algorithms learn that competitors will immediately match price cuts, they'll soon learn to keep prices high. In 2011, duelling Amazon Marketplace algorithms responding to each other's price changes caused an out-of-print genetics book to be listed at $23.6 million.[19]

Price escalation and lock-in may become commonplace, even without criminal intent.

Evidence collapse

Two lynchpins of contemporary evidence – speech and video – will soon be falsifiable, further distorting the persuasive landscape. Convincing text-to-speech software already exists, although it's computationally hungry, and undetectable 'deepfake' videos are just a couple of years away. When we can simply import incriminating text and see an accurate rendition from our most hated politician, we will have to reconsider what we believe to be true. Photos are already useless as evidence: once audio and video follow suit, what can serve as an accurate record of fact?

Audio synthesis firm Lyrebird is among the few tech companies to publish an ethics statement.

> Imagine that we had decided not to release this technology at all. Others would develop it and who knows if their intentions would be as sincere as ours: they could, for example, only sell the technology to a specific company or an ill-intentioned organization. By contrast, we are making the technology available to anyone and we are introducing it incrementally so that society can adapt to it, leverage its positive aspects for good, while preventing potentially negative applications.[20]

That the statement exists is welcome, but the contents are lousy. (We'll discuss its flimsy ethical argument – 'If we don't, someone else will' – in chapter 5.) It's not enough for transformative technology companies to warn of ethical risk and leave society to figure it out. This is instrumentalism at its most dangerous; technologists must actively understand and mitigate the harms their products can do.

Evidence collapse threatens not only our understanding of facts and current affairs, but also our personal relationships. If we can't be sure whether the person on the phone or video call is who we think it is, the door is open for widespread manipulation. Trust-based technologies like cryptography or blockchain might help, but these will require excellent, consumer-grade design. If only a few techies can

implement these safeguards, information anarchy will reign for everyone else.

Justifying persuasion: folk ethics

Clearly, persuasion has complex implications. We have to ask that evergreen ethical question: where should we draw the line? What divides the unethical dark pattern from beneficent persuasion? Let's start with some common ethical precepts.

The weakest justification for persuasion – or, indeed, anything else – is that everyone's at it. This is a classic ethical trap, identified centuries ago by David Hume and known as the *is-ought fallacy*. It's a theoretical error to derive what we should do (ought) from how people currently act (is). Our competitors' moral choices are irrelevant to our own. Just as a cheating peloton didn't excuse Lance Armstrong's drug use, OkCupid's blasé defence of persuasive experimentation as commonplace stumbles straight into the is-ought jaws.

The *golden rule* – do as you would be done by – is more helpful. This proverb of reciprocity is found in ancient belief systems from Leviticus to Confucius. Applied to persuasive design, the golden rule suggests we should only persuade someone to do something we'd do ourselves, or that we'd be happy for someone to persuade us of. The golden rule's biggest flaw is its egocentrism. It encourages everyone to see themselves as the ideal ethical arbiter, whether their interests align with others' or not. The golden rule ignores the variety of human desires and the role of context in ethical choices.

Perhaps we should instead treat others how *they* would like to be treated: the *platinum rule*. In other words, we should only persuade people to act in their own interests. We should pause here to distinguish individual interest from the public interest, an idea often found in journalistic ethics. Stories that operate in grey ethical areas, such as those that infringe privacy or involve deception, often undergo a public interest test. This decision weighs up potential harm to individuals against the wellbeing of society; an editor-in-chief will often deem stories that increase accountability and transparency at the expense of wrongdoers to be in the public good.

Some persuasive technologies have a public interest component. A weight-loss app could prevent thousands of obesity-related deaths.

But the common good is politically charged: attempts to specify how others should live often bear the taint of authoritarianism. The public interest is complex, and not always a helpful ethical focus.

If we should only persuade people to act in their own interests, who gets to decide what those interests are? Technologists probably aren't the right people for this decision; they typically favour the scientific over the spiritual, action over reflection, and progress over the status quo, values which may not be right for the individual in question. But simply asking people what their best interests are has its flaws too: people's stated opinions are unreliable, and at times everyone contradicts their own interests by seeking out things that limit their capacity to thrive, like tobacco and alcohol.

If we can't just ask people what their best interests are, and it's improper to specify interests on others' behalf, we're in a bind, torn between a paternalistic desire to help others and a tolerant respect for people's freedom to choose. These simple ethical guidelines – forgive the pejorative label, 'folk ethics' – don't solve the problem.

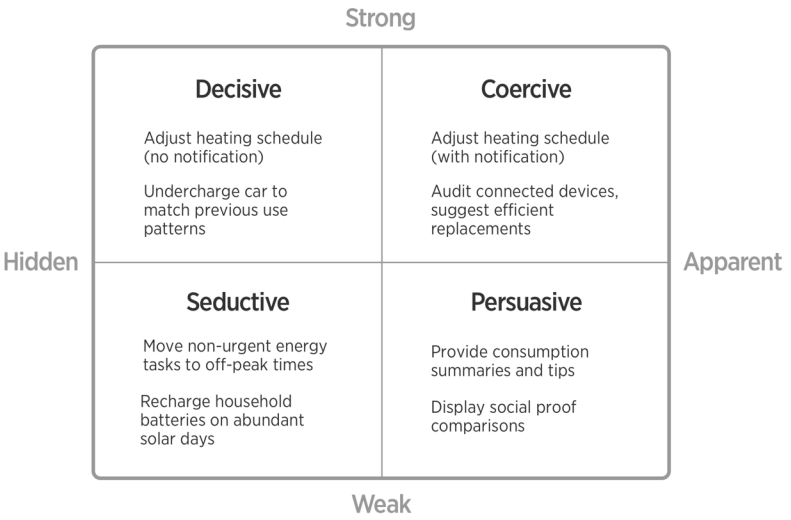
Persuasive theories

A few designers and scholars have proposed guidelines for persuasive systems. Daniel Berdichevsky and Erik Neuenschwander suggest, among other principles, we should judge persuasion on whether it would be appropriate in person, freed from technology.[21] This question has the valuable side effect of restoring the personal context technology so often strips away. In *Persuasive Technology*, BJ Fogg suggests that using negative emotions to persuade is ethically questionable.[22] Mass persuasion does plenty of this – advertisers play on envy; bitter politicians appeal to rank xenophobia – but we should be wary of the is-ought fallacy. Warnings on cigarette packets are perhaps more defensible; there's a case that the ends (saving lives) justify the means (using fear to persuade). If, however, we do reject the idea of tugging at the subject's negative emotions, we should prevent technologies from displaying these emotions too. An AI should never shout at its user for skipping an upgrade, no matter how improved the new firmware may be.

Almost all theorists agree it's unethical to mislead for persuasive purposes, including Richard Thaler, who includes it in his principles

of ethical nudging.[23] But consider the *placebo button*, a functionless control such as the close door button in many elevators, or the save option in certain web apps. The motivation – giving users a sense of control – is benevolent, the means deceptive. Are placebo buttons unethical? They still respect the user's will – the door still closes, the settings are still stored – and perhaps a white lie is preferable to the truth: you aren't in control, the technology is. However, if in doubt, deceptive persuasion is best avoided.

Social design researcher Nynke Tromp suggests we classify persuasion by strength and visibility, creating four types of influence: decisive, coercive, persuasive, and seductive.[24] Let's say we're designing a smart energy hub and want people to conserve energy. Here are some potential design approaches, mapped to these four categories.



The ethical implications of the top-left quadrant seem the most significant. A cold house might be dangerous to the elderly, and an undercharged vehicle may prove disastrous in an emergency, but if the device takes unilateral, invisible decisions both could happen.

Strong forms of persuasion may at times be justified, but weaker forms usually place us on safer ethical ground. Luciano Floridi distinguishes informational nudges from structural nudges.[25] An informational nudge changes the nature of information available – labelling

unhealthy snacks, for example – while a structural nudge changes the courses of actions available, such as moving these same snacks out of easy reach. The informational nudge is weaker than the structural nudge, but more respectful of free choice.

Persuasive objects in the physical world are usually visible or even highlighted, such as speed cameras, but digital constraints are often invisible. With a software volume lock in place, users will never know just how much louder and more harmful their headphones could be. Is disclosure the answer, then? Should persuasive technologies simply advertise their presence and methods? This is a promising idea, with two caveats. First, persuasion may require invisibility; disclosing persuasion might make it ineffective. Second, disclosing every single persuasive method would be messy and distracting. Explanations and warnings would litter our technologies; users would eventually just start to ignore them. Many persuasive techniques are just part of what we consider good design, such as ensuring labels are clear and calls to action are highlighted. There has to be some balance. Disclosing persuasive methods is a noble aim, but perhaps it's better to make this information available rather than prominent. We'll discuss an example shortly.

The role of intent

We should also consider disclosing our persuasive intent – the why behind the design. Intent comes up often in ethics, and is the cornerstone of the *principle of double effect*, which states that harm is sometimes acceptable as a side effect of doing good. Double effect is often used in euthanasia cases: doctors may increase a dying patient's morphine dose to ease suffering, even in the knowledge the dose may prove fatal. If the intent were simply to kill the patient, the doctor would be morally and legally liable, but most authorities choose not to prosecute when a convincing double-effect defence (relieving pain) exists.

Most experts suggest ethical persuasion needs positive intent. In Thaler's words, 'there should always be a good and clear reason for why the nudge will improve the welfare of those being nudged.' We should be honest about our true intent when designing persuasive systems. Why do I want people to follow my advice? What's in it for

them? What's in it for me? Unfortunately, intent is less useful when examining other people's choices. 'Did you mean well?' is vulnerable to all sorts of excuses; as Benjamin Franklin observed, being 'a reasonable creature [...] enables one to find or make a reason for every thing one has a mind to do.' An unscrupulous colleague can concoct a plausible upside for virtually any ethical transgression. The overpriced extended warranty? Invaluable to the small fraction of users whose product breaks. Sucking up the user's contacts without permission? Imagine how thrilled people will be when they learn their friends have joined! Suggest someone acted with impure intent and they'll often respond with twisted double-effect arguments and who-me gestures of mock indignation. Focusing just on intent also allows us to wriggle off the hook of unintended consequences. Users don't care whether we intend harm or not; they care whether we cause harm.

Introducing deontology

Persuasive theories and honest questions about intent are useful ethical tools, but perhaps we need something more rigorous, some set of moral rules to follow. This is the foundation of *deontological ethics* (or duty ethics), one of the three schools of modern ethics. Deontologists believe that ethics is governed by rules and principles, and that we have a moral duty to adhere to these rules. This can make deontologists somewhat rigid: if we believe we have a moral duty to always tell the truth, it's hard to justify lying to the secret police about where our family is hiding. Deontologists lead lives of principle, but also lives of self-denial and, occasionally, honourable suffering. That said, deontologists typically excel at resisting ethical pressure; their belief in rules and integrity mean they set clear boundaries and challenge bad behaviour.

Immanuel Kant, a pioneer of deontological thought, proposed a powerful idea: when faced with an ethical choice, we should universalise our thinking. Kant suggested we imagine whether our actions would be acceptable as a universal law of behaviour. *What if everyone did what I'm about to do?* This simplified version of Kant's most important theory[26] is an invaluable ethical prompt for technologists. It

focuses us on the futures our decisions could create and forces us to see ethical choices from broader social perspectives.

Kant also posed another useful deontological question: *am I treating people as ends or means?*[27] This deserves some explanation. For our purposes, the question asks whether we're using people – users, stakeholders, wider society – for our own success, or treating them as autonomous individuals with their own goals. Designers usually don't struggle with the ends-or-means question, since they tend to believe deeply in the importance of users' goals. The question tends to be more difficult when we ask it about company-wide decisions, particularly those that affect millions of people.

While deontologists agree we should live according to moral rules, they don't specify those rules: the point is we have to figure them out as a society. The questions above are good prompts, but we still have to work hard to translate them into action. Let's see how our two ethical tests help us untangle our persuasive complications.

Should we ship a deceptive dark pattern that offers no user benefit but increases our profit? Well, what if everyone did what I'm about to do? If all technology were riddled with dark patterns, companies may earn more, but our technologies – and probably our lives – would be worse. Users would feel hoodwinked, and we'd squander the trust our industry urgently needs. So the deontological answer is clear: no, we shouldn't release this dark pattern.

How about the attention economy? A world in which we all paid for our beloved products with attention rather than cash wouldn't of itself be bad; although, as we'll soon see, there may be painful privacy implications. The problems arise when a user is truly addicted, to the point that it harms their overall wellbeing. If we encourage addicts to use our services, are we treating these people as ends or means? That's easy: means. We continually offer them something that harms them, while we profit. A deontologist will argue tech companies have a duty to intervene in cases of harmful use. Unlike a tobacco company, who can't cut off a specific smoker, tech companies could identify problem users from afar and take action. This could involve anything from a light touch – reducing notifications or showing a 'Time for a break?' fatigue alert – to total excommunication, banning a user's credit card and closing their account. A company that know-

ingly serves an addicted user is using that addict as a means for commercial success alone, and is crossing the ethical line.

Ethical experimentation

If we put experiments under the deontological microscope, there's plenty to improve. Our first deontological test – what if everyone did what I'm about to do? – suggests the idea of running experiments isn't itself too harmful; it's our methods that cause the problems.

First, users have no choice about whether to take part. Usually, every user can be co-opted into an experimental population; however, mandatory research doesn't allow for informed consent. This decision clearly reduces people's autonomy, and would make a bad universal law of behaviour. Second, experiments are opaque. People usually have no way to know which experimental groups they're in and when the tests will end. Opacity can't be a healthy universal principle either. Third, in some companies, the point of experimenting becomes not improving the product but hitting targets: teams throw out different approaches until people respond in the right way. This is the very definition of treating people as means, not ends.

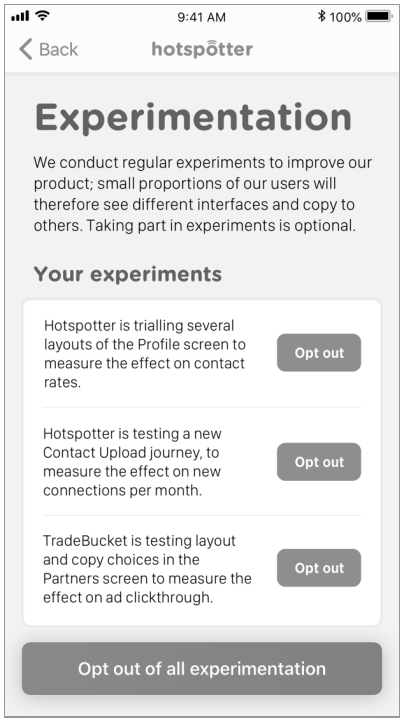
Can we design a more ethical, deontological approach to experimentation, one we'd recommend as a universal method? Let's start by always considering users as ends, not means. We should pledge that we'll try to improve the user experience with every experiment, and decline to run experiments we believe will be neutral or harmful. Businesses often need to take decisions that users won't like, such as raising prices or pruning functionality; under this principle these changes aren't suited to experiments. If you want to raise prices, raise prices across the board. Pledging to always improve users' experiences would fix many of the problems with Facebook's emotional contagion study: researchers would have to ensure participants only saw happier updates, not more negative ones.

In recognition that experimentation is research, we should consider informed consent inviolable. Since children can't give consent, we should remove them from the experimental pool, unless we can also gain consent from guardians. We should also agree that users should be able to learn about experiments they're in, with some screen or notification that describes each experiment, tells the user

who's running the research, and describes the goal, phrased perhaps in terms of the metrics we're tracking or the hypotheses we're testing.

Let's also draw from persuasive theory and pledge to avoid both negative emotion and deception in our experiments, and strive to use weak, visible forms of persuasion where possible, such as informational nudges rather than structural nudges. Finally, users should be able to opt out of individual experiments and the whole experimentation programme alike, with no negative effects. Users who opt out will still receive software updates once they're rolled out to all users; we just won't include these people in experiments.

For a hypothetical smartphone app, a single screen could satisfy many of these requirements:



This more ethical approach to experimentation wouldn't be too onerous. We'd have to be more rigorous in selecting sample populations, add a new screen or notification, provide short metadata for each experiment, and build reliable opt-in and opt-out systems. These

features needn't be highly prominent in the product, so long as they're available and findable. These small changes should help us treat users with respect, reduce the risk of regulatory wrath, and mean we run experiments in a way we'd be happy for others to follow.

The veil of ignorance

Another useful principle for designing fair systems is John Rawls's *veil of ignorance*. In *A Theory of Justice*,[28] Rawls contends that society – for our purposes we'll extend this to technological society too – is best structured as if its architects didn't know their eventual role in the system. Beneath a veil of ignorance, we wouldn't know our social status, our intelligence, or even our interests; but if the system is fair we should be satisfied wherever we ended up.

The veil of ignorance has some links with deontology; you wouldn't want to emerge from behind the veil into a role in which you're just a means for everyone else's ends. You may also recognise some parallels with the golden rule. However, this isn't just about treating people how you'd like to be treated yourself, but creating entire systems in which everyone is treated fairly. It's the you-cut-I-choose cake-sharing protocol from your childhood, stretched over an entire population.

Rawls's idea, focused on equality and redistributive justice, attracts criticism from predictable quarters when applied to politics, but for our purposes it's powerful. The veil of ignorance forces us to consider all the various roles people will play in our systems, and how our work might influence people from wide-ranging backgrounds. Applied to persuasion, the veil of ignorance suggests we should only create persuasive systems that would be fair to the persuader and persuaded alike.

Better persuasion

Some of our persuasive difficulties are direct consequences of our product development processes. The strategies that create desirable products also foster addiction. We use our phones 150 times a day; is that because they're designed to capture our attention, or because we

genuinely love them? Probably both. That we find it difficult to tease these motivations apart speaks to the failures of the experience design movement. Designers haven't interrogated the difference between enjoyable and habitual use, and the rhetoric of designing for delight has directly contributed to addiction.

Some commentators argue that to counter manipulation, tech firms should actively expose people to conflicting views. The tech industry has plenty of tools at its disposal: fact-checking plug-ins and trust ratings to combat disinformation, and crowdsourcing, blockchains, and cryptography to fight evidence collapse. But counteracting people's biases is tough, thankless work. Early efforts to diversify the information environment have been exasperatingly crude: Facebook's attempts to warn users of suspect content actually made more people click on it, and I've only recently escaped an infuriating experiment that appended the most engaging reply (usually the most contentious or trollish) to every News Feed article.

Isn't this just more technocratic meddling, though? The idea that technologists should force-feed the masses balanced information diets should trouble us. Which harm is more severe: the threat of manipulation, or the authoritarian threat of controlling others' information environments? These questions are social, political, and legal as much as they are technical; as such, they aren't for us to answer alone. The technical fixes that would be most effective against disinformation, such as real name policies or better tracking of sources, would themselves endanger privacy. Perhaps our most important duty is to stimulate public discussion about persuasive technologies. The tech industry should look to boost information literacy at all levels of education and adult life, and play an active role in restoring a thriving, resilient press. Technologists may need to give users anti-addiction and anti-persuasion strategies, or even build counter-technologies that side with the user against the industry itself, such as persuasion blockers that scrub out manipulative advertising and bust people out of non-consensual A/B tests.

We should also eliminate the factors that have caused our persuasive woes. At the heart of the dark pattern, the addictive app, and the disinformation problem alike lies an undue fixation on quantification and engagement. Choosing new success metrics would smooth the route to more ethical persuasion. The Time Well Spent movement[29]

asks how tech would look if it were designed to respect human values rather than capture attention. The movement taps into theories of calm technology and mindfulness to inspire designers to protect users' time and agency, and argues for new business models that subvert the attention economy.

Quantitative data should always be paired with accessible qualitative research, so human stories can claim their rightful place in decision-makers' minds. We can also select *mutually destructive targets*, metrics chosen in pairs such that one will suffer if we simply game the other. For example, dark patterns may well extract more revenue per user, but they'll also harm retention if users feel duped. Choosing both revenue and retention as mutually destructive targets provides a minor safeguard against abuse; if both measures move in the right direction, we can be confident things are genuinely improving.

Regulation and opt-out

If the industry fails to self-police, it should brace itself for consumer rejection. Until recently, society saw technological refuseniks as socially irregular, and it was mostly techies themselves who chose to abstain, deleting their apps, climbing mountains, and writing think-pieces about their experiences. These efforts reeked of privilege – after all, you need to be rich to need nothing – but amid growing concern about addictive technologies, a public temperance movement is brewing. Clinics are already treating self-described app addicts; perhaps a detox-as-a-service industry will emerge: hand over your devices and we'll lock you out for two weeks.

Where consumers lead, regulators will follow. Tech companies have already been sued and subpoenaed over unfair persuasion and dark patterns; in 2015, LinkedIn paid $13 million to settle a dark pattern class action suit. Regulation is likely to come first from the EU, given its historical opposition to tech monopolies and its citizens' sensitivity to corporate abuse. The German government is drafting a law that would impose €50 million fines on social networks that fail to curtail hate speech and disinformation. Regulators might decide to make platforms liable for hate speech, force tech conglomerates to split, or demand that social networks let users take their friend networks to competitor services. Online adverts might, and arguably

should, be required to reveal their funders. Some philosophers and lawyers are even discussing whether there should be an enshrined legal right to attentional protection.

The early television age also spawned concerns about persuasion and disinformation. Many governments responded by establishing national broadcasting agencies and standards, creating a heavy top-down influence on the burgeoning industry. If this pattern is repeated for emerging persuasive technologies, the industry will have only itself to blame.

Thanks for reading this free sample of Chapters 1, 2, and 3 of Future Ethics.

Buy the full book (9 chapters, paperback, PDF, Apple Books, and Kindle formats):

future-ethics.com

NOTES

1. Trouble in paradise

1. Richard Sennett, *The Craftsman* (Penguin, 2009).
2. 2017 Cone Communications CSR Study, conecomm.com.
3. See Bruno Latour, *Pandora's Hope: Essays on the Reality of Science Studies* (Harvard University Press, 1999) for one analysis of this argument.
4. Peter-Paul Verbeek, *Moralizing Technology* (University of Chicago Press, 2011). Verbeek in turn draws on the work of Don Ihde and Latour.
5. Melvin Kranzberg, 'Software for Human Hardware?', in Pranas Zunde & Dan Hocking (eds.), *Empirical Foundations of Information and Software Science V* (Plenum Press 1990).
6. One of those seductive quotes of hazy origin, often attributed to Maxim Gorky, and spoken in Jean-Luc Godard's film *Le petit soldat*, attributed to Lenin.
7. Caroline Whitbeck, *Ethics in Engineering Practice and Research* (Cambridge University Press, 2nd ed., 2011).

2. Do no harm?

1. Or so says economist Horst Siebert, at least. The tale is possibly apocryphal, but even an anecdote can contain an undeniable truth: things don't always turn out as planned.
2. Paul Virilio, *Politics of the Very Worst* (Semiotexte, 1999).
3. Don Ihde, *Technology and the Lifeworld* (Indiana University Press, 1990).
4. Shannon Hall, 'Exxon Knew about Climate Change almost 40 years ago', *Scientific American*, 26 Oct 2015, scientificamerican.com.
5. See Thomas Wendt, 'Decentering Design or a Critique of Human-Centered Design', slideshare.net.
6. Ben Thompson, 'Airbnb Versus Hotels', *Stratechery*, 18 Apr 2017, stratechery.com.
7. Ursula Franklin, *The Real World of Technology* (House of Anansi Press, 2nd ed., 1999).
8. David Ingold and Spencer Soper, 'Amazon Doesn't Consider the Race of Its Customers. Should It?', *Bloomberg*, 21 Apr 2016, bloomberg.com.
9. Joanna Bryson, 'Three very different sources of bias in AI, and how to fix them', 13 Jul 2017, joanna-bryson.blogspot.com.
10. Aylin Caliskan, Joanna Bryson, Arvind Narayanan, 'Semantics derived automatically from language corpora contain human-like biases', *Science*, 14 Apr 2017, 183–186.
11. Chris Ip, 'In 2017, society started taking AI bias seriously', *Engadget*, 21 Dec 2017, engadget.com.
12. Andrew Thompson, 'Google's Sentiment Analyzer Thinks Being Gay Is Bad', *VICE Motherboard*, 25 Oct 2017, vice.com.
13. Laura Hudson, 'Technology Is Biased Too. How Do We Fix It?', *FiveThirtyEight*, 20 Jul 2017, fivethirtyeight.com.

14. Stella Lowry and Gordon Macpherson, 'A blot on the profession', *British Medical Journal* Vol. 296, 5 March 1988.
15. Christian Rudder, 'Race and Attraction, 2009–2014', *OkCupid*, 10 Sep 2014, okcupid.com.
16. Lizzie Edmonds, 'Google forced to remove vile racist search suggestions from its site for a number of British cities including Bradford, Leicester and Birmingham', *MailOnline*, 11 Feb 2014, dailymail.co.uk.
17. 'I think it's time we broke for lunch…', *The Economist*, 14 Apr 2011, economist.com.
18. Carlota Perez, *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages* (Edward Elgar Publishing, 2003).
19. *Terminator 2: Judgement Day*, dir. James Cameron (TriStar Pictures, 1991). You could argue these aren't Sarah's words but John's, passed on by Kyle through Sarah back to John.
20. Originally conceived by military strategist Charles Taylor and adapted by several futurists since, including Joseph Voros.
21. Octavia Butler, 'A Few Rules for Predicting the Future', *Essence Magazine*, 2000.
22. Shannon Vallor, *Technology and the Virtues* (Oxford University Press, 2016).
23. Genevieve Bell, 'Rage Against the Machine?', talk at *Interaction12* conference.
24. Eric Meyer and Sara Wachter-Boettcher, *Design for Real Life* (A Book Apart, 2016).
25. My name, extending an idea by Sam Jeffers.
26. Cameron Tonkinwise, 'Ethics by Design, or the Ethos of Things', *Design Philosophy Papers*, 2:2, 129-144, 2004.
27. Jared Spool, 'Creating Great Design Principles: 6 Counter-intuitive Tests', *UIE*, 1 Mar 2011, uie.com.
28. Anand Giridharadas, 'A tale of two Americas. And the mini-mart where they collided', talk at *TED2015* conference, ted.com.
29. Shannon Vallor, 'An Introduction to Data Ethics: a resource for data science courses', *Markkula Center for Applied Ethics*, scu.edu.

3. Persuasive mechanisms

1. Langdon Winner, 'Do Artifacts Have Politics?', *Daedalus* 109, no. 1 (1980): 121-36.
2. Daniel Berdichevsky and Erik Neuenschwander, 'Toward an ethics of persuasive technology', *Communications of the ACM*, 42, 5 (May 1999), 51–58.
3. Coined by Harry Brignull; see darkpatterns.org.
4. Tim Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (Knopf Publishing Group, 2016).
5. Peter Kafka, 'Amazon? HBO? Netflix thinks its real competitor is... sleep', *Recode*, 17 Apr 2017, recode.net.
6. Kleiner Perkins Internet Trends 2017, kpcb.com/internet-trends.
7. Ian Bogost, 'The App That Does Nothing', *The Atlantic*, 9 Jun 2017, theatlantic.com. Bogost is a theorist of persuasive games and designer of *Cow Clicker*, a notorious commentary on the manipulative aspects of FarmVille. Cow Clicker players had but one objective: to click a cow. The game amassed 50,000 users before Bogost triggered a 'Cowpocalypse', vanishing the cows into ironic rapture.
8. Adam Cramer, Jamie Guillory, Jeffrey Hancock, 'Experimental evidence of massive-scale emotional contagion through social networks', *Proceedings of the National Academy of Sciences (PNAS)*, 17 Jun 2014 vol. 111 no. 24 8,788–8,790.
9. Ariel Rubinstein and Ayala Arad, 'The People's Perspective on Libertarian-Paternalistic Policies' (2015).

10. Molly Sauter, 'The Apophenic Machine', *Real Life Magazine*, reallifemag.com.
11. Chris Wade, 'The Reddit Reckoning', *Slate*, 15 Apr 2014, slate.com.
12. A remarkable inconsistency discovered by journalist Olivia Solon.
13. Berit Anderson and Brett Horvath, 'The Rise of the Weaponized AI Propaganda Machine', scout.ai.
14. Philip Howard, and Bence Kollanyi, 'Bots, #Strongerin, and #Brexit: Computational Propaganda during the UK-EU Referendum.' Working Paper 2016.1. Oxford, UK: Project on Computational Propaganda.
15. Jeff Bezos, '2015 Letter to Shareholders'.
16. Paul Robinette et al., 'Overtrust of robots in emergency evacuation scenarios', *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, 2016, pp. 101-108.
17. Sandra Matz et al., 'Psychological targeting as an effective approach to digital mass persuasion', *Proceedings of the National Academy of Sciences (PNAS)* 2017.
18. Dana Mattioli, 'On Orbitz, Mac Users Steered to Pricier Hotels', *The Wall Street Journal*, 23 Aug 2012, wsj.com.
19. John D. Sutter, 'Amazon seller lists book at $23,698,655.93 – plus shipping', *CNN*, 25 Apr 2011, cnn.com.
20. Lyrebird, 'With great innovation comes great responsibility', lyrebird.ai/ethics.
21. Daniel Berdichevsky and Erik Neuenschwander, 'Toward an ethics of persuasive technology', *Communications of the ACM*, 42, 5 (May 1999), 51–58.
22. BJ Fogg, *Persuasive Technology* (Morgan Kaufman, 2003).
23. Richard Thaler, 'The Power of Nudges, for Good and Bad', *New York Times*, 31 Oct 2015, nytimes.com.
24. Nynke Tromp et al., 'Design for Socially Responsible Behavior', *Design Issues*, Volume 27, Number 3, Summer 2011.
25. Luciano Floridi, 'Tolerant Paternalism: Pro-ethical Design as a Resolution of the Dilemma of Toleration', *Science and Engineering Ethics* (2016), 22: 1669.
26. His first formulation of the 'categorical imperative', from *Groundwork for the Metaphysics of Morals*.
27. Again from the categorical imperative, this time the second formulation.
28. John Rawls, *A Theory of Justice* (Harvard University Press, 1971).
29. Now run by the Center for Humane Technology, humanetech.com.

ABOUT THE AUTHOR

Cennydd Bowles is a London-based designer and writer with fifteen years of experience and clients including Twitter, Ford, Cisco, and the BBC. His focus today is the ethics of emerging technology. He has lectured on the topic at Carnegie Mellon University, Google, and New York's School of Visual Arts, and is a sought-after speaker at technology and design events worldwide.

Also by Cennydd: *Undercover User Experience Design* (New Riders 2010), with James Box.