

The Psychology of Technology

How Algorithms Change
What We Think, Want, and Do

*An Interdisciplinary Monograph on Cognition,
Motivation, and Behaviour in Algorithmically
Mediated Environments*

Contents

Preface	xvii
The Argument in One Picture	xxi
I Foundations of Algorithmic Psychology	1
1 The Algorithmic Turn	3
1.1 An inversion of adaptation	3
1.2 Why psychology cannot treat this as someone else's problem	5
1.3 Three verbs: think, want, do	7
1.4 The loop, stated precisely	8
1.5 What this book claims, and what it does not	10
1.6 A note on evidence and its grades	11
1.7 The road ahead	12
1.8 Four objections, and why they do not dissolve the problem	13
1.9 What would falsify the argument	14
Chapter summary	15
2 Inside the Machine	17
2.1 Why the technical detail matters	17
2.2 The anatomy of a ranking system	17
2.3 Prediction from digital traces: how much can be inferred . .	21
2.4 Generative models as a new kind of interface	22
2.5 What is and is not knowable from outside	23
2.6 A worked example: the anatomy of one impression	23
2.7 Why fixing this is harder than it looks	24
Chapter summary	25
3 The Human Substrate	27
3.1 Designing for a mind that already exists	27
3.2 Attention is a scarce, contested resource	27

3.3	Two systems, unequally engaged	28
3.4	Reward, prediction, and the value of uncertainty	29
3.5	Habit: the automation of the self	30
3.6	The social mind: comparison, belonging, and reputation . .	31
3.7	Self-determination and the conditions of motivation	31
3.8	Preferences are constructed, not retrieved	32
3.9	Sources of resistance	33
3.10	Individual differences: who is most exposed	33
3.11	What psychology gets wrong when it studies this badly . .	34
	Chapter summary	35
 II Changing What We Think		37
4	The Attention Economy	39
4.1	The commodity nobody agreed to sell	39
4.2	Capture: the mechanics of attention acquisition	40
4.3	Fragmentation and its cognitive consequences	41
4.4	Measurement problems that distort the whole field	42
4.5	From aggregate hours to specific mechanisms	43
4.6	Attention as a collective, not merely individual, resource . .	43
4.7	The economics of an externality	44
4.8	Attention as a developmental and clinical variable	45
4.9	Practical implications for attention	45
	Chapter summary	46
5	The Outsourced Mind	47
5.1	Offloading is old; its scope is not	47
5.2	Search engines and the reorganisation of memory	48
5.3	Generative assistants: offloading inference, not just storage	49
5.4	Calibration: the real variable of interest	50
5.5	What the individual gains, the collective may lose	51
5.6	A practical framework for delegation	52
5.7	The education case	52
5.8	Metacognitive calibration and the fluency trap	53
5.9	What a healthy division of cognitive labour would look like	54
	Chapter summary	54

6	The Manufacture of Belief	57
6.1	A field that overcorrected twice	57
6.2	The machinery of belief formation	57
6.3	The empirical correction: what audits actually found	60
6.4	Reconciling mechanism and measurement	62
6.5	What actually reduces the spread of falsehood	63
6.6	Why corrections are weak and prevention is strong	65
6.7	The generative turn in the epistemic environment	65
6.8	Collective knowledge and machine-mediated culture	66
	Chapter summary	67
III	Shaping What We Want	69
7	The Machinery of Desire	71
7.1	Wanting without liking	71
7.2	Prediction error and the value of uncertainty	72
7.3	Social reward as the reinforcer	73
7.4	From engagement to problematic use	74
7.5	Motivation that supports well-being, and motivation that does not	75
7.6	What happens when the reinforcement is withdrawn	75
7.7	Designing reward differently	76
7.8	Gamification and the quantification of everything	76
7.9	Boredom, and what it was for	77
7.10	Case study: the anatomy of a streak	78
	Chapter summary	78
8	Constructed Preferences	81
8.1	The circularity at the heart of personalisation	81
8.2	How malleable are preferences, really?	82
8.3	Social proof and the manufacture of popularity	83
8.4	Preference amplification and the incentive to change you . .	83
8.5	Does narrower mean worse?	85
8.6	Generative personalisation: the next turn	85
8.7	Diversity as a design objective	86
8.8	The regress problem in preference elicitation	87
8.9	Markets, monoculture, and correlated error	88
	Chapter summary	88

9	The Mirror and the Metric	91
9.1	The most contested question in the field	91
9.2	Mechanisms: why one would expect an effect	91
9.3	What the evidence shows	93
9.4	Identity under algorithmic conditions	95
9.5	What a careful reading of this literature yields	96
9.6	Design features rather than platforms	97
9.7	The self as an object of measurement	98
	Chapter summary	98
IV	Directing What We Do	101
10	Choice Architecture at Scale	103
10.1	From cafeteria to interface	103
10.2	How large are nudge effects?	104
10.3	Digital nudging and its instruments	105
10.4	Dark patterns: when architecture becomes deception	106
10.5	Personalised nudging: the sharpest case	107
10.6	Nudging versus boosting	108
10.7	Case study: the consent interface	109
10.8	Personalisation of interventions: promise and hazard	110
10.9	Self-nudging: choice architecture turned inward	110
	Chapter summary	111
11	From Habit to Compulsion	113
11.1	Automaticity as the design objective	113
11.2	Where habit shades into compulsion	114
11.3	Self-control: what does and does not work	115
11.4	Digital self-control tools and their modest record	116
11.5	Why individual solutions are insufficient	116
11.6	The measurement of automaticity	117
11.7	Withdrawal, substitution, and relapse	118
11.8	A protocol that follows from the evidence	118
	Chapter summary	119
12	Deciding with Machines	121
12.1	Two failure modes, opposite in direction	121
12.2	When people accept machine advice	122
12.3	Why explanations do not fix it	123
12.4	Does the combination help at all?	124

12.5 Algorithmic management: when the machine decides about you	125
12.6 Design principles for machine-assisted judgement	126
12.7 Expertise, deskilling, and the long horizon	126
12.8 Trust, its calibration, and its social dimension	127
12.9 Autonomy at work under algorithmic supervision	128
Chapter summary	128
V Ethics, Autonomy, and the Future	131
13 Manipulation and Autonomy	133
13.1 Everything influences; that cannot be the objection	133
13.2 A conception of autonomy that survives	134
13.3 The ethics of nudging, revisited digitally	135
13.4 Consent, privacy, and their limits	137
13.5 Fairness, distribution, and who bears the cost	137
13.6 Whose values? The alignment question	138
13.7 Against both technological determinism and technological innocence	139
13.8 Four cases, graded	140
13.9 Responsibility without a responsible party	141
13.10 The paternalism objection, taken seriously	142
Chapter summary	142
14 Governing the Loop	145
14.1 Four levers, unequally powerful	145
14.2 Measurement: the highest-leverage change	145
14.3 Design: what to build	146
14.4 Regulation: what the evidence supports	148
14.5 A research agenda	149
14.6 What determines the outcome	150
14.7 What each actor can do	151
14.8 Three scenarios	152
14.9 The discipline this requires	153
Chapter summary	153
Epilogue: The Choosing Animal	155
Glossary	159

CONTENTS

Bibliography	163
Index	189

CHAPTER 1

The Algorithmic Turn

The environment is no longer a stage on which behaviour occurs. It is a player.

The question of this chapter: what changes for psychology when the environment a person inhabits is itself an adaptive system optimising against a model of that person?

1.1 An inversion of adaptation

Consider two descriptions of the same activity, separated by twenty years. In the first, a person sits at a desktop computer, opens a web browser, types a query into a search box, scans a list of blue links, and clicks one. The system is passive. It waits, matches keywords, and returns. Whatever the person ends up reading is the product of their query, their scanning strategy, and a ranking function that knows almost nothing about them personally.

In the second, a person picks up a phone in a queue. Before any explicit intention has formed, an application opens to a stream of items that no other person on earth will see in that order. The selection reflects several thousand implicit signals: what this person watched to completion last Tuesday, how long they hovered before scrolling past a particular item, what people whose behavioural embeddings sit near theirs engaged with in the last four hours, the time of day, the device, the connection speed, and the system's current estimate of which item will maximise the probability of a session extending past some threshold. The person did not issue a query. They did not specify a topic. The system produced the intention as much as it served one.

The difference between these two scenes is not merely one of degree. It is a change in the causal structure of the situation. In the first, the environment is fixed with respect to the person, and psychology can proceed

in its customary way: hold the environment constant, vary something about the person or the task, and measure the response. In the second, the environment is a function of the person, and the person is a function of the environment, and the two are updated continuously against each other. This is a closed loop, and closed loops have properties that open ones do not. They can amplify small perturbations. They can converge on states that no participant chose. And they can render certain measurements circular: if a system's estimate of what you want determines what you are offered, and what you are offered determines what you choose, then choice data is no longer an independent test of the estimate.

Key idea

Classical psychology largely assumes an environment that does not optimise against the participant. Contemporary digital environments violate that assumption systematically. The violation is not a nuisance parameter; it changes the interpretation of core constructs, including preference, attitude, and choice.

The phrase *algorithmic turn* names this shift. It is not a claim that algorithms are new—sorting and ranking procedures are as old as bureaucracy—nor that personalisation is new—Gillespie (2014) showed that the very notion of algorithmic “relevance” encodes a set of contestable public commitments about what knowledge matters and to whom. It is a claim about the conjunction of four properties that became simultaneously true only recently: environments are now *instrumented* (nearly every micro-behaviour is logged), *predictive* (models forecast individual behaviour with non-trivial accuracy), *individuated* (the environment differs across people rather than across coarse segments), and *closed-loop* (the environment updates on the consequences of its own choices, often within hours). Any one of these alone is manageable. Together they underwrite what Zuboff (2015) named surveillance capitalism: an economic logic in which behavioural surplus—data exhaust beyond what is needed to deliver the service—is claimed as raw material for prediction products sold in behavioural futures markets, with the accumulation of behavioural modification capability as its terminal asset (Zuboff, 2019). One need not accept every element of that account to recognise the structure it describes. Together these four properties constitute a genuinely novel object of psychological study, and one that Rahwan et al. (2019) argued requires its own empirical discipline—a science of *machine behaviour* that studies algorithmic systems the way ethologists

study animals, through their behaviour in the environments they actually inhabit rather than through their source code alone.

1.2 Why psychology cannot treat this as someone else's problem

There is a tempting division of labour in which computer scientists study the algorithms, sociologists study the institutions, lawyers study the regulation, and psychologists continue to study the mind as though its environment were incidental. The division fails for three reasons.

1.2.1 The mechanisms being exploited are psychological

The design levers that make contemporary digital products effective are not novel inventions of engineering. They are direct applications of well-characterised findings from the psychology of learning, attention, and social cognition. Intermittent, uncertain reward delivery sustains behaviour more robustly than predictable reward—an effect established in operant conditioning long before there were notification badges, and grounded neurally in the finding that midbrain dopamine neurons encode a reward *prediction error* rather than reward itself (Schultz et al., 1997; Schultz, 2016). Defaults are extraordinarily powerful determinants of choice, a result demonstrated most memorably in organ-donation registration rates across European countries that differ in whether consent is opt-in or opt-out (Johnson and Goldstein, 2003). Repeated exposure increases both liking (Zajonc, 1968) and perceived truth (Fazio et al., 2015). Judgement under uncertainty runs on heuristics whose systematic deviations from normative standards were catalogued half a century ago and are now exploitable at scale (Tversky and Kahneman, 1974). Social comparison is an automatic and consequential process (Festinger, 1954). People economise on effortful cognition wherever an external store will do the work (Risko and Gilbert, 2016; Sparrow et al., 2011).

The design discipline that assembled these levers named itself explicitly: Fogg (2009) formalised a behaviour model in which action occurs when motivation, ability, and a trigger converge, and prescribed that designers who cannot raise motivation should instead raise ability by removing effort—which is precisely the friction-minimisation strategy analysed in Chapter 11.

A designer who arranges variable-ratio social feedback, opt-out defaults, high-frequency repetition, salient comparison metrics, and one-tap offloading is not doing something exotic. They are operating a well-stocked labora-

tory of twentieth-century psychology at population scale, with continuous outcome measurement, and without the constraints of informed consent that govern the discipline that produced the findings. It would be strange for psychology to have supplied the mechanisms and then declined to study their deployment.

1.2.2 The measurements have changed

For most of its history, psychology worked with small samples, coarse temporal resolution, and self-report. Algorithmic environments generate behavioural traces at millisecond resolution for hundreds of millions of people, and permit randomised experiments of a scale the discipline could previously only imagine—experiments involving 61 million participants in a single study of political mobilisation (Bond et al., 2012), or 689,003 participants in a study of emotional contagion whose publication triggered a decade-long ethical argument (Kramer et al., 2014). This is an extraordinary methodological opportunity and a serious hazard. The opportunity is obvious. The hazard is subtler: platform-scale behavioural data measures what was *done inside a designed environment*, not what was wanted, and the two can diverge arbitrarily. A meta-analytic literature now documents that self-reported and log-derived measures of digital media use correlate only moderately—in the range typically reported around $r \approx 0.4$ —which means that a substantial fraction of the published literature relating “screen time” to outcomes is measuring something other than what it claims (Parry et al., 2021). Wagner et al. (2021) make the broader case: measuring an *algorithmically infused society* requires instruments that account for the fact that the measurement apparatus and the phenomenon are entangled.

1.2.3 The stakes are psychological

Finally, the outcomes at issue are the discipline’s own dependent variables. Whether adolescents’ well-being has been affected by social platforms is a psychological question that has been argued with unusual heat and unusual methodological sophistication (Twenge et al., 2018; Orben and Przybylski, 2019a; Odgers and Jensen, 2020; Valkenburg et al., 2022; Vuorre and Przybylski, 2024). Whether exposure to algorithmically ranked political content changes attitudes is a psychological question, and one that has now been addressed with platform-scale randomised trials (Guess et al., 2023a,b; Nyhan et al., 2023). Whether reliance on generative models degrades or augments reasoning is a psychological question currently being answered in real time (Lee et al., 2025; Gerlich, 2025; Vaccaro et al., 2024). If psychology

does not answer these questions with its own methodological standards, they will be answered by advocacy.

1.3 Three verbs: think, want, do

The organising device of this book is a decomposition of algorithmic influence into three levels, shown in Figure 1.1. The levels are distinguished by what is changed, over what timescale, and by which psychological machinery.

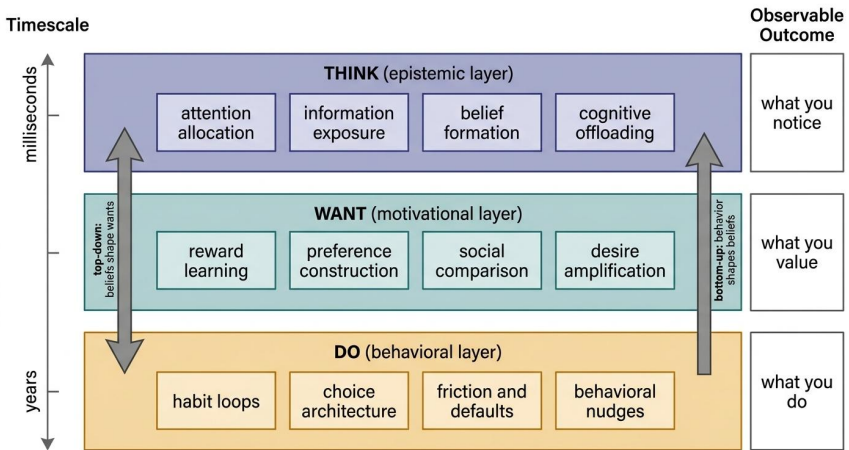


Figure 1.1: Three levels of algorithmic influence. The epistemic layer (THINK) concerns attention and belief; the motivational layer (WANT) concerns reward learning and preference; the behavioural layer (DO) concerns habit and choice. The layers are coupled in both directions and operate on timescales spanning milliseconds to years.

Think. At the epistemic layer, algorithmic systems determine the informational environment: which items are seen, in what order, how often, and with what framing. Because attention is the scarcest resource in an information-rich world—the point Herbert Simon made in 1971 and which remains the founding observation of the field—selection is not a neutral service but the exercise of consequential power. Repeated selection produces familiarity, familiarity produces fluency, and fluency is systematically misread as evidence of truth (Fazio et al., 2015; Pennycook et al., 2018). Selection also determines what is *not* seen, and the psychology of absent evidence is unforgiving: people do not spontaneously correct for information they never encountered.

Want. At the motivational layer, algorithmic systems shape the schedule and character of reward. Contemporary products deliver social and informational rewards on variable schedules with short latencies, engaging reward-prediction machinery that evolved under very different statistics (Berridge and Kringelbach, 2015; Lindström et al., 2021). Over longer horizons, recommendation systems participate in the construction of taste itself. Preferences are not stable objects that systems merely discover; the constructive view of preference has been standard in judgement and decision-making research for three decades (Slovic, 1995; Bettman et al., 1998; Ariely and Norton, 2008). A system that learns and then reinforces an inferred preference is intervening in an ongoing construction process, not reading out a fixed parameter.

Do. At the behavioural layer, algorithmic systems arrange choice architecture: defaults, friction, salience, sequencing, and timing. The behavioural-science literature established that such arrangements move behaviour (Thaler and Sunstein, 2003), though the magnitude is contested—published nudge effects appear inflated by publication bias, and the largest field evaluations report considerably smaller effects than the academic literature suggests (Mertens et al., 2022; DellaVigna and Linos, 2022). What is distinctive digitally is not the potency of any single nudge but the ability to select, personalise, and iterate nudges continuously against measured response (Weinmann et al., 2016; Caraban and Karapanos, 2020; Yeung, 2017). Over repetition in stable contexts, behaviour becomes automatic, and automaticity is the defining property of habit (Wood and Neal, 2007; Wood and Rünger, 2016; Lally et al., 2010).

The three layers are not independent. Beliefs constrain what is wanted; wants direct attention; repeated behaviour generates beliefs about the self. The arrows in Figure 1.1 run both ways, and much of the interesting dynamics lives in the coupling rather than in any single layer.

1.4 The loop, stated precisely

Figure 1.2 states the mechanism this book examines. It is worth walking through its five stages carefully, because most confusions about algorithmic influence are confusions about which stage is doing the work.

Stage 1, behaviour. A person clicks, dwells, scrolls, skips, likes, purchases, abandons. These are the raw observables.

Stage 2, instrumentation. Behaviour is converted into events with identities, timestamps, and contexts. What is not instrumented is invisible

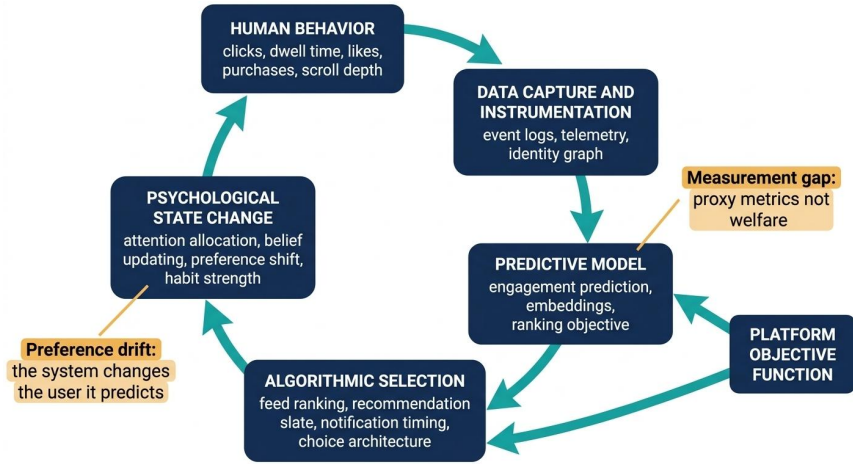


Figure 1.2: The algorithmic feedback loop. Behaviour is instrumented, modelled, and used to select the next environment; the selection changes the person, who generates the next behaviour. Two failure modes recur throughout this book: the objective is a proxy for value rather than value itself, and the system’s own outputs contaminate the data from which it learns.

to everything downstream, and what is invisible cannot be optimised for. Regret, boredom, and the quiet satisfaction of a session that ended at the right moment are rarely instrumented.

Stage 3, prediction. A model learns to forecast a target variable from these events. The target is chosen by the organisation. Almost always it is a proxy: predicted click-through, predicted watch time, predicted probability of returning tomorrow.

Stage 4, selection. A policy converts predictions into an environment—a ranked feed, a recommendation slate, a notification schedule, a set of defaults and prices.

Stage 5, psychological change. The selected environment alters attention allocation, belief, preference, and habit strength. The altered person then returns to Stage 1.

Two structural pathologies follow from this diagram, and both recur in every subsequent chapter.

The first is the **proxy problem**. The quantity optimised is not the quantity valued. Engagement is easy to measure and correlates with value in the short run; it also correlates with outrage, novelty, and compulsive

checking. There is now direct evidence that optimising engagement proxies amplifies content that users themselves, on reflection, do not endorse or benefit from (Stray et al., 2024a; Milano et al., 2020). Content that provokes out-group animosity is disproportionately shared (Rathje et al., 2021); negativity increases consumption of news (Robertson et al., 2023); morally and emotionally charged material spreads further within ideological networks (Brady et al., 2017, 2020). None of this requires malice. It follows from optimising a proxy in a population whose attention is differentially captured by threat and moral emotion.

The second is **feedback contamination**. Because the system’s own selections generate the data from which it subsequently learns, the training distribution is not a sample from the world but a sample from the world as the system has already shaped it. This is not a speculative concern. Chaney et al. (2018) demonstrated in simulation that algorithmic confounding of this kind increases homogeneity of consumption and, notably, *decreases* the utility users derive from the system; Jiang et al. (2019) formalised the conditions under which such loops degenerate; and Kalimeris et al. (2021) and Kasirzadeh and Evans (2023) showed that recommenders trained by reinforcement learning face structural incentives to *shift* user preferences toward more predictable states, because a more predictable user is easier to score well on. The system’s accuracy improves partly by making the user more like its model of them.

Key idea

An adaptive system that is rewarded for predicting a person has a latent incentive to make that person more predictable. Preference change is not a side effect of recommendation; under a range of standard objectives it is instrumentally useful to the optimiser.

1.5 What this book claims, and what it does not

Because this field is prone to overstatement, it is worth fixing the scope of the argument precisely at the outset.

The book claims that algorithmic environments exert real, measurable influence on attention, belief, preference, and behaviour; that the mechanisms of that influence are identifiable and drawn from well-established psychology; that the closed-loop structure of these environments creates dynamics—preference drift, homogenisation, proxy divergence—that are absent from classical accounts of media influence; and that these dynam-

ics raise questions about autonomy that are not adequately addressed by existing notions of consent.

The book does not claim that algorithms determine what people think. The empirical record is emphatic on this point. Large-scale randomised interventions on platform ranking algorithms during a United States election campaign found substantial changes to what users *saw* and comparatively small or null changes to political attitudes and polarisation over the study period (Guess et al., 2023a,b; Nyhan et al., 2023). Ideological segregation online is real but concentrated in a small minority of heavy consumers (Flaxman et al., 2016; Hosseinmardi et al., 2021), and the average person’s information diet remains more diverse than the popular “filter bubble” story implies (Bakshy et al., 2015). Exposure to opposing views does not reliably reduce polarisation and in at least one well-powered field experiment increased it (Bail et al., 2018). Any account that treats algorithmic influence as straightforwardly hypodermic is not merely simplistic; it is contradicted by the strongest available evidence.

The interesting position lies between these poles, and it is a specifically psychological position. Algorithmic systems rarely change what a well-informed adult believes about a salient political question over a three-month window. They routinely change what is available to think about, which questions feel urgent, how effortful it is to verify a claim, how often a comparison is made, and what the default action is. Those are not small things. They are the conditions under which cognition happens, and psychology has known since its behaviourist beginnings that conditions of that kind do most of the causal work in everyday life.

1.6 A note on evidence and its grades

Throughout the book, claims are graded by the strength of the design behind them. It will help to state the hierarchy explicitly.

1. **Platform-scale randomised field experiments** with preregistration and independent analysis are the strongest available evidence on the effects of algorithmic systems, and there are still remarkably few of them (Guess et al., 2023a; Allcott et al., 2020; Bond et al., 2012).
2. **Individual-level randomised experiments** outside platforms—deactivation studies, abstinence trials, intervention tests—are strong on internal validity and often weak on ecological validity and duration (Hunt et al., 2018; Tromholt, 2016; Allcott et al., 2022).

3. **Algorithmic audits**, in which researchers instrument sock-puppet or donated accounts to characterise what systems actually serve, establish exposure but not effect (Ribeiro et al., 2020; Hosseinmardi et al., 2021; Huszár et al., 2022).
4. **Longitudinal and intensive-repeated-measures designs** can identify within-person dynamics and heterogeneity that group averages conceal, and have repeatedly shown that averaged effects hide large individual differences in sign as well as magnitude (Beyens et al., 2020; Valkenburg and Peter, 2013).
5. **Cross-sectional correlations** between self-reported use and outcomes are the weakest common design and, as specification-curve analyses have shown, are highly sensitive to analytic choices (Orben and Przybylski, 2019a; Ferguson et al., 2025).

A recurring pattern in this field is that the strongest designs produce the smallest average effects, while the weakest designs produce the headlines. The pattern is not proof that the effects are illusory. Average effects can be small while distributional effects are severe; population-level consequences can accumulate from per-exposure effects too small to detect in any single trial (Bak-Coleman et al., 2021); and the outcomes most likely to matter—the slow reshaping of what one finds interesting, the erosion of the capacity for sustained attention—are the hardest to capture in a three-week randomised trial. But the pattern does impose an obligation of care, and this book tries to meet it.

1.7 The road ahead

Chapter 2 opens the machine. It explains, without mathematics beyond arithmetic, what a recommender system is, what an embedding represents, how ranking objectives are chosen, why A/B testing produces a particular kind of institutional myopia, and how systems come to optimise things nobody intended. Chapter 3 does the same for the human being, assembling the components of the mind that these systems engage: limited attention, dual-process cognition, reward prediction, habit, social motivation, and constructed preference.

With both halves of the loop specified, Parts II through IV trace the consequences, and Part V asks what should be done. Figure 1.3 provides orientation: the technologies discussed in this book have a history, and the psychological findings they exploit are in most cases older than the technologies themselves.

From Broadcast to Algorithmic Curation: A Brief History

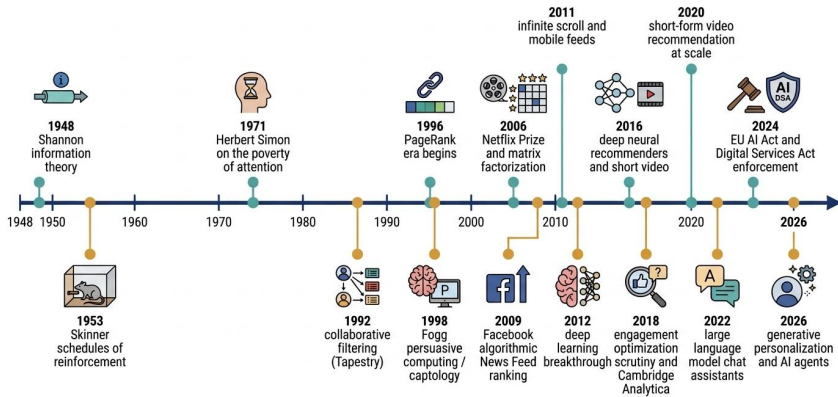


Figure 1.3: From broadcast to algorithmic curation. The psychological principles exploited by contemporary systems substantially predate them; what changed was the capacity to apply those principles individually, continuously, and at scale.

1.8 Four objections, and why they do not dissolve the problem

A book making the claims above should meet the strongest objections early rather than accumulate them in a reader’s mind through fourteen chapters. Four are worth stating in their most forceful form.

“This is the same moral panic that greeted every previous medium.” The observation is correct and the inference from it is not. Novels, newspapers, cinema, radio, comic books, and television each provoked confident predictions of cognitive and moral collapse that did not materialise, and the base rate for such predictions is accordingly poor. But the argument from base rates is only as good as the reference class, and the four properties identified above—instrumentation, prediction, individuation, and closure—were absent from every earlier medium. A television broadcast did not observe which viewers looked away, estimate the probability that each would return tomorrow, and recompose itself accordingly for each household. The correct response to the moral-panic objection is not to dismiss it but to insist on mechanism: the claims in this book are tied to identifiable processes with measurable signatures, and each is falsifiable.

“The measured effects are tiny.” Often they are, and this book says so repeatedly. Three considerations prevent the observation from settling the matter. First, small per-exposure effects applied continuously to billions of people are not small in aggregate; the relevant quantity is the integral, not the instantaneous value. Second, average effects conceal distributions, and several chapters here document heterogeneity in which a minority experiences effects an order of magnitude larger than the mean (Beyens et al., 2020; Valkenburg et al., 2022). Third, the outcomes most plausibly affected—what a person finds interesting, how long they can attend, what they take for granted—are precisely those that short experiments cannot measure.

“People are not passive; they use these systems deliberately.”

Entirely true, and the literature on folk theories and algorithmic resistance in Chapter 3 takes it seriously (Eslami et al., 2015; Karizat et al., 2021; Swart, 2021). Agency and influence are not mutually exclusive. A person can be a sophisticated, strategic, resistant user of a system that nonetheless shapes their attention, and the demonstration that users theorise about algorithms does not show that their theories are accurate or that acting on them is effective.

“Preferences have always been socially constructed, so nothing has changed.”

This is the most serious objection and it is answered throughout Part III. The short form: what is new is not that preferences are shaped by environments but that the environment is now an optimiser with a measurable objective, an individually fitted model of the person, and a feedback channel that lets it evaluate whether its shaping worked. Social construction by an indifferent culture and social construction by a system with a loss function are different processes, and only the second has a gradient.

1.9 What would falsify the argument

A book that cannot be wrong is not a scientific book. The claims made here imply specific observations that would count against them, and it is worth recording them.

If systematic manipulation of ranking objectives produced no detectable change in consumption diversity, taste trajectories, or reported satisfaction over multi-year horizons, the preference-amplification argument of Chapter 8 would be substantially weakened. If the demand for commitment devices

documented by Allcott et al. (2022) disappeared when measured with better instruments, the habit-and-self-control account of Chapter 11 would lose its principal empirical support. If well-powered studies found that explanations reliably improved discrimination rather than merely trust, the pessimistic reading of transparency in Chapter 12 would be wrong (Bansal et al., 2021). And if interventions built on the mechanisms described here—accuracy prompts, friction, reflective elicitation—consistently failed to move the outcomes the mechanisms predict, the mechanistic story would be in trouble regardless of how coherent it sounds.

None of these has occurred so far. Several have not been properly tested, which is itself a finding, and Chapter 14 treats it as the core of the research agenda.

Chapter summary

- Digital environments have become instrumented, predictive, individuated, and closed-loop. The conjunction of these four properties, not any one alone, constitutes the algorithmic turn.
- The design mechanisms these environments use are applications of established psychology: variable reinforcement, defaults, repetition, social comparison, and cognitive offloading.
- Two structural pathologies follow from the loop’s architecture: optimisation targets are proxies that diverge from value, and the system’s own outputs contaminate the data from which it learns, creating incentives for preference change.
- The strongest available evidence—platform-scale randomised trials—finds large effects on *exposure* and modest effects on *attitudes* over short horizons. Claims of direct algorithmic determination of belief are not supported.
- The correct level of analysis is conditional rather than hypodermic: algorithmic systems change the conditions under which thinking, wanting, and acting occur.

