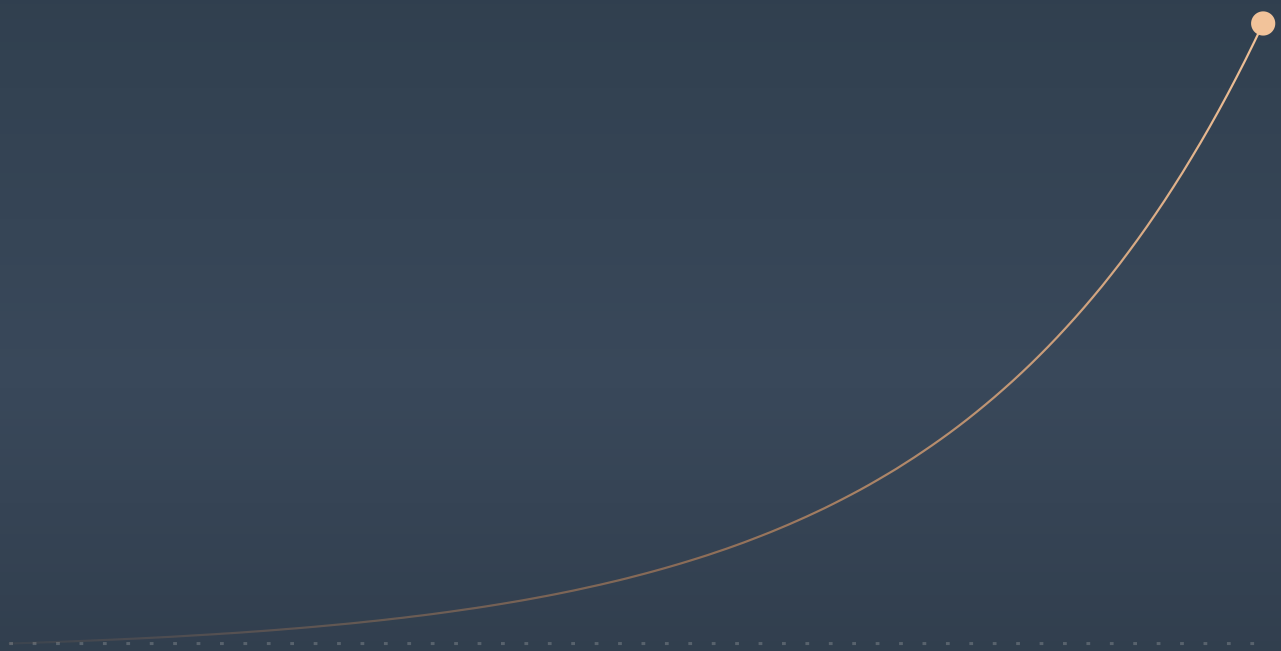
A FIELD GUIDE TO THE HIDDEN COST OF PERSUASIVE DESIGN

# Behavioral Debt

What persuasive design really costs, and how to build products that do not borrow against trust.



**Amogh Tiwari**

WORKING MANUSCRIPT · DRAFT EDITION · 2026

THE ARGUMENT IN ONE SENTENCE

*Persuasive software works by borrowing against the user's trust. The debt is invisible because no dashboard has a line for it. It compounds until a product is most hated at the exact moment it looks most engaged. The only real repayment is a design that gives the user back what was taken.*

CONTENTS

# What is in the book

Three chapters written to publication quality, the full taxonomy, six original frameworks, and nine pattern entries. Every subsection below is a link. The remaining chapters of the trade edition are planned but not yet drafted.

REFERENCE AND APPARATUS



BACK MATTER

METHOD

# The constitution

The rules every chapter, pattern entry and figure is written against. The Citation Law comes first because it is the one failure that cannot be repaired after submission.

### The Psychology of Consumer Software

A SYSTEMS-LEVEL REFERENCE ON BEHAVIORAL DESIGN PATTERNS

This file is the constitution of the project. Every chapter, pattern entry, teardown, and figure is written against these rules. When in doubt, this file wins.

---

## 1. What this is

A research document that answers one question:

**How does consumer software change human behavior, and by what mechanisms?**

It is built as a **taxonomy-first** work. The taxonomy of behavioral patterns is the spine; chapters, teardowns, and frameworks all hang off it. We do not write prose that is not anchored to a pattern, a framework, or a cited finding.

**Primary deliverable:** a course report with a fixed page target and deadline. **Secondary:** the same source files extend into a longer handbook without rewriting.

Scope is cut to fit the deadline. When the two conflict, the deadline wins and we ship a narrower document that is *complete and defensible* rather than a broad one that is thin.

---

## 2. THE CITATION LAW (non-negotiable)

This is the rule that can sink the project, so it comes before everything else.

An AI can generate a reference list that looks immaculate, plausible authors, real-sounding journal names, well-formed DOIs, specific page numbers, and is entirely fabricated. It will survive every read-through. It will not survive a professor checking one entry. That is an academic-integrity failure, not a typo.

Therefore:

1. **No citation is written unless the source was actually retrieved and read.**

Not "recalled," not "well known," not "obviously by Kahneman." Retrieved.

1. **Every empirical or factual claim is in exactly one of two states:**
2. **Grounded**, traceable to a source in bibliography/references.md, which records the

URL/DOI actually fetched and the date fetched.

- **Marked**, carries a visible CITATION NEEDED tag in the text.

There is no third state. A confident, uncited, unmarked factual assertion is a bug.

1. **Never invent:** DOIs, page numbers, publication years, sample sizes, effect sizes,

participant counts, quotations, or study results. If the specific number is not in front of us, we write the claim without the number or we mark it.

1. **Numbers get special scrutiny.** "Mathur et al. surveyed 11,000 shopping sites" is exactly

the kind of sentence that is 90% right and 10% fabricated. Any figure like this must come from the retrieved source, and the reference entry must say where.

1. **Original frameworks are exempt and must be labeled.** The Behavioral Influence Model,

Behavioral Genome, Behavioral Cost Index, Behavioral Debt, Pattern DNA, and the Behavioral Design Pyramid are the author's own synthesis. Write them freely, but never dress them as established findings, and never cite a fake source in support of them. Label: "*Original framework; proposed here.*"

1. **Separate fact from interpretation.** A cited finding and our reading of it are different

sentences. Do not blend them into one.

1. **Replication status is part of the citation.** Behavioral science has a replication crisis and

this document is about behavioral science. Any claim resting on a contested effect must say so in the text. This is a *feature*: a report that says "the effect everyone cites here does not replicate, and here is the one that does" is stronger than one that repeats the folklore.

Already caught by verification, and load-bearing:

- **"Fogg's equation B = MAT" does not exist.** The 2009 paper was retrieved and read in full.

It contains **no equation**. The mnemonic is a community retrofit. Fogg also renamed "Trigger" to "Prompt" in late 2017 (now B=MAP). Our Behavioral Influence Model is *inspired by* Fogg, not derived from him, and must say so.

- **The Zeigarnik effect does not replicate** (2025 meta-analysis: pooled recall ratio 0.99).

The effect that does is **Ovsiankina** (task *resumption*, 67%). Progress bars and unfinished onboarding are resumption phenomena. Citing Zeigarnik for them is near-universal in UX writing and is wrong.

- **Never print "19% vs 34%"** for the Nunes & Drèze car-wash study. Widely repeated; not

confirmed against the primary text. The 8-vs-10-step manipulation *is* confirmed.

- **Never cite a "nudge effect size."** The pooled effect is disputed under publication-bias

correction. Cite Nudge as a framework, or cite specific studies.

- **Cite Ferster & Skinner (1957), not "Skinner."** Dropping Ferster is the most common

citation error in persuasive-design writing.

**Consequence to accept honestly:** targets like "300+ references" and "100+ papers reviewed" are now real work with a real cost, not numbers that can be hit by generation. If the deadline does not permit 300 verified sources, the document says so and cites fewer. A short honest bibliography beats a long fictional one.

#### ENVIRONMENT LIMITS (STATE THESE, DO NOT PAPER OVER THEM)

- pandoc is **not installed**. The PDF pipeline is an open toolchain decision, deferred.
- **"1000+ screenshots" is not achievable** from this environment. Figures are Mermaid/text

diagrams plus a small number of manually captured screenshots. Do not imply otherwise.

---

### 3. The three goals every section must satisfy

A section that fails any of these gets cut or rewritten.

| GOAL | TEST |
|---|---|
| **Academic rigor** | Claims are grounded or marked. Sources are primary where possible. |
| **Industry relevance** | A named, real product exhibits the thing being described. |
| **Original contribution** | It says something the literature does not already say. |

If a passage is only a summary of existing work, it belongs in the literature review, not in a chapter.

---

### 4. The valence rule

We are **not** writing "dark patterns are bad." We are mapping a complete behavioral system in which the same mechanism can be benign or harmful depending on whose interest it serves.

Every pattern is scored on a three-point **valence**:

- **Ethical**, user's interest and business interest align; user retains autonomy.
- **Persuasive**, business interest leads; user autonomy is preserved; disclosure is honest.
- **Manipulative**, works *because* the user does not understand what is happening to them,

or would object if they did.

The valence is a property of the *implementation*, not the *mechanism*. A streak is ethical in a language app the user wants to stick with; it is manipulative when it manufactures guilt to prevent a user from leaving. Say which, and why.

This is what makes the analysis defensible rather than moralizing.

---

### 5. Taxonomy structure

Three axes. This is the original contribution and the spine of the document.

- **Family** (mechanism), A through I, see TAXONOMY.md
- **Lifecycle stage** (where it acts), Discovery → Install → Registration → Activation →

Onboarding → Trust → Habit → Routine → Retention → Monetization → Expansion → Exit

- **Valence** (whose interest), Ethical / Persuasive / Manipulative

Every pattern carries a stable ID (F-##) that never changes once assigned. Cross-references use the ID, never the title, so titles can be revised without breaking the document.

---

### 6. Pattern entry template

Every pattern in patterns/ uses exactly this structure. No exceptions, consistency across 100+ entries is what makes it a reference work rather than a collection of essays.

```
# [ID], Pattern Name

**Family:** | **Lifecycle stage:** | **Valence:** | **Severity:** /5

## Definition

One paragraph. What it is, stated so precisely that a reader could identify it in the wild.

## Pattern DNA

Goal → Trigger → Interface → Bias → Emotion → Decision → Behavior → Reward → Business metric

## Psychological basis

The mechanism. Cited, or marked [CITATION NEEDED].

## Interface expression

How it actually shows up on a screen.

## Real-world examples

Named products. What specifically they do. Observed or sourced, never assumed.

## Business rationale

The metric it moves and why a rational company ships it.

## Behavioral Cost Index

Time / Attention / Memory / Trust / Money / Privacy / Emotion, scored, with justification.

## Valence analysis

Why this instance is ethical, persuasive, or manipulative. The honest version.

## Ethical alternative

A design that achieves the legitimate business goal without the harm. Concrete, not a platitude.

## Regulatory status

Any law that touches it. Cited, or "no known regulation."

## Related patterns

By ID.

## Sources
```

---

## 7. App teardown template

Every app in apps/ uses this structure.

```
# App Name

**Category:** | **Business model:** | **Primary metric (inferred):**

## Behavioral genome

Attention / Habit / Identity / Trust / Autonomy, scored 0-10, with justification per axis.

## Pattern inventory

Table: Pattern ID | Where it appears | Valence as implemented | Severity

## Behavioral flow

The user's path through the app, stage by stage, annotated with the patterns acting at each.

## What it does well

Mandatory section. If we cannot fill it, we have not understood the product.

## Where it crosses the line

Specific, evidenced, and separated from opinion.

## Sources

Screenshots (dated), public reporting, regulatory filings, the app itself.
```

**Evidence rule for teardowns:** describe only what has actually been observed in the product or is documented in a retrievable source. Do not describe a screen from memory. Products change; memory of a 2023 interface is not evidence about the 2026 product. Date every observation.

---

### 8. Writing rules

- Academic register. No marketing voice. No hype.
- Prefer primary sources (the paper, the regulation, the filing) over secondary (the blog post about the paper).
- Every chapter ends with: **Key Takeaways**, **Open Questions**, **References**.
- Define a term the first time it is used; never assume the reader shares our vocabulary.
- Short sentences carry technical weight better than long ones.
- Tables for enumerable facts. Prose for reasoning. Never reason inside a table cell.
- No em dashes in headings or callouts.
- Figures are Mermaid where possible so they live in version control as text.

CHAPTER 01

# How Software Changes What You Do

**The question this chapter answers:** *How does software change human behavior?*

*It answers that question and no other. It does not yet ask whether the change is good, whether it is legal, or what should be done about it. Those are Chapters 2 through 9. This chapter establishes only the mechanism, because every normative argument later in the document depends on the mechanism being described correctly first.*

---

### 1.1 Software is not a tool, it is a behavioral environment

The default folk theory of software is that it is a tool. A tool is an object with a purpose. It sits inert until a person picks it up, and when they put it down it stops acting on them. A hammer does not have an opinion about how often you use it. It has no stake in whether you reach for it tomorrow. The relationship is asymmetric in the user's favour: the tool is passive, the user is the agent, and all the intention in the transaction belongs to the person holding it.

This theory is not merely incomplete for consumer software. It is inverted.

Consumer software is not inert between uses. It sends. It schedules. It waits for the hour at which you historically opened it and it arrives then. It maintains state about you that persists across sessions and shapes what you will be shown when you return. It has a measurable interest in your return, quantified, dashboarded, and owned by a named person whose compensation depends on it. The intention in the transaction is not unilaterally the user's. Some of it belongs to the product, and the product's portion of that intention was designed, tested, and deployed with more deliberation than the user brings to the encounter.

The precise claim is this: **a piece of consumer software is not an object a person acts upon. It is an environment a person acts inside.** And environments are not neutral. They have gradients. Some paths through them are downhill and some are uphill, and which is which was decided by someone.

This is not a metaphor that has been stretched for rhetorical effect. It has an empirical consequence that is testable, and it has been tested. Luguri and Strahilevitz (2021) placed representative samples of American consumers in front of an offer for a questionable data-protection service and varied nothing about the offer itself. The service was the same. The price was the same. What changed was the interface: the sequence of screens, the default framing, the wording of the decline option, the obstruction placed in front of refusal. In the control condition, where users could simply accept or decline on the first screen, 11.3 percent accepted. Under what the authors classified as mild dark

patterns, 25.8 percent accepted. Under aggressive dark patterns, 41.9 percent accepted. Study 1 had a final sample of 1,963 participants; Study 2, which decomposed the effect by strategy, had 3,777.

Their own summary of the result is the sentence this entire document is built on top of:

"*Decision architecture, not price, drove consumer purchasing decisions.*" *(Luguri & Strahilevitz, 2021)*

Read that as an empirical finding, not as a slogan. Price is the variable that economics puts at the centre of the consumer decision. In this experiment, holding the product and the price fixed and varying only the environment moved acceptance from roughly one in nine to roughly two in five. The environment was not a modifier on the decision. In the authors' data, the environment substantially *was* the decision.

If a change in the arrangement of screens can do that to a purchase, then the arrangement of screens is not a presentational layer sitting on top of the real product. It is a causal input to human behavior, of the same kind and roughly the same order as the thing being sold. Design is not the packaging around the decision. Design is a term in the decision.

The rest of this chapter develops what follows from taking that seriously.

---

### 1.2 Why consumer software matters

A methodological confession belongs at the start of this section rather than buried in its footnotes.

The intuitive way to open a chapter like this is with scale. Cite the number of smartphone users. Cite the average daily screen time. Cite the count of applications installed on a median device, or the number of notifications received per day. These figures are readily available in aggregator reports and they are repeated so often that they feel like common knowledge.

This document does not print them, because the Citation Law of this project (PROJECT.md, section 2) prohibits writing a number that has not been retrieved from a primary source and recorded with a fetch date, and no such source has been retrieved for any of them. The temptation to write "the average person checks their phone 96 times a day" is exactly the failure mode the Citation Law exists to prevent: a figure that is 90 percent plausible, endlessly repeated, and traceable to nothing. Accordingly:

- Global smartphone penetration figures: CITATION NEEDED
- Average daily screen time: CITATION NEEDED
- Average notifications received per day: CITATION NEEDED
- Number of applications installed on a median device: CITATION NEEDED
- Time spent in social, video, or messaging applications specifically: CITATION NEEDED

These markers are not decorative. They are an instruction to a future editor: obtain the primary source or delete the sentence.

What can be argued without any of those numbers is the *structural* case, and the structural case is in fact the stronger one, because it does not depend on any particular year's usage statistics remaining true.

**First: the surface is universal in kind, whatever its size.** The claim is not that software touches a specific fraction of human waking hours. The claim is that the categories it mediates are the categories that structure a life. Communication with other people. Access to money. Access to health information. Access to news. Access to romantic partners. Access to employment. Access to transport. Each of these was, within living memory, mediated by an institution with a physical location, a human counterparty, and a regulatory regime built around that physicality. Each is now mediated, for a very large number of people, by an interface. Whether that number is 60 percent or 90 percent of the population does not change the argument. The argument is about *what kind of thing* is now being decided inside interfaces, not how many people are inside them.

**Second: the decision surface is where the choice is now made.** A person choosing a mortgage in 1990 made that choice in a room, with a document, in the presence of a person whose obligations to them were legally specified. A person choosing a subscription in 2026 makes that choice on a screen, in a flow whose sequence was chosen by the seller, whose default was chosen by the seller, and whose decline option was worded by the seller. The migration of the decision into the interface is a migration of the decision into an environment the counterparty controls end to end.

**Third: iteration.** A physical retail environment can be redesigned perhaps once a decade. A software environment can be redesigned continuously, and, critically, it can be redesigned *against measured evidence of what makes users comply*. This is the point Luguri and Strahilevitz make about the mechanism of proliferation, and it is worth quoting because it is the causal engine of everything in this document: "Dark patterns are presumably proliferating because firms' proprietary A-B testing has revealed them to be profit maximizing." The interface is not merely a controlled environment. It is a controlled environment with a feedback loop attached, optimising against a behavioral objective, at a cadence no physical environment can match.

**Fourth: asymmetry of attention.** The product's side of the encounter is the output of a professional team, with quantitative feedback, working over months. The user's side is a few seconds of divided attention. There is no version of this encounter in which the two parties are bringing comparable resources to bear. This asymmetry is not a moral accusation. It is a description of the situation, and it holds even when everyone involved is acting in good faith. It is the reason a chapter about mechanism has to come before any chapter about ethics: the asymmetry is structural, and it exists whether or not it is exploited.

The argument of this section, then, is not "software is everywhere, therefore it matters." It is: **the decisions that most determine a person's material and social life are now taken inside environments whose gradients are set by a counterparty, tuned against measured behavioral response, and iterated faster than any regulator, competitor, or user can adapt.** That is true regardless of what the screen-time figure turns out to be.

---

## 1.3 The environment argument

The claim that an environment has behavioral gradients is not new to software. It is old, and it was made most clearly by people studying physical retail and physical gambling. The value of the analogy is not decorative; it is that both those environments were understood as behavioral machines long before software was, and the vocabulary developed for them transfers cleanly.

#### THE SUPERMARKET

Nothing about a supermarket's layout is accidental, and nothing about it is hidden. The staples are placed at the far end of the store so that the path to them crosses the discretionary aisles. Eye-level shelving is the most valuable shelving and is allocated accordingly. The checkout lane is stocked with small, cheap, high-margin, impulse-priced goods positioned at the exact moment when the shopper is stationary, has already committed to a purchase, and is no longer in a deliberative frame.

What is instructive is that none of this is deception. There is no lie anywhere in a supermarket. Every price is printed. Every item can be put back. A shopper who wanted to buy only milk could walk directly to the milk and directly back out. The supermarket does not prevent that shopper. It simply makes that path slightly uphill and every other path slightly downhill, and then relies on the fact that across a very large number of shoppers, a small gradient produces a large aggregate effect.

**The supermarket does not defeat the deliberate shopper. It harvests the inattentive one.** This distinction is the single most useful idea in the analogy, and it recurs in every family of the taxonomy. Almost no behavioral pattern in consumer software works by overpowering a person who is paying full attention. Nearly all of them work by being placed where full attention is not being paid.

Thaler and Sunstein gave this idea its canonical name. Their publisher credits them, in their own words, with inventing the term **choice architecture**: the observation that there is no neutral way to present a choice, that some arrangement must be picked, and that whatever is picked will influence the outcome (Thaler & Sunstein, 2008; 2021). The lack of a neutral option is the load-bearing half of the claim. A designer cannot decline to have an effect on behavior. A designer can only decline to *notice* the effect they are having.

A caveat required by this project's replication clause: the *Nudge* framework is cited here as a framework and as a source for the term "choice architecture," which is verified. It is not cited as a source for any effect size. The nudge literature has a contested pooled effect under correction for publication bias, and no effect size for "nudges in general" appears anywhere in this document.

Thaler and Sunstein's later contribution is more directly useful to us. In the 2021 Final Edition they introduce **sludge**: friction deliberately placed between a person and what they actually want. Sludge is the authors' own term for the adversarial inverse of a nudge, and it is, for our purposes, the conceptual bridge between benign choice architecture and the manipulative end of the spectrum described in section 1.10. Sludge is what a cancellation maze is made of.

#### THE CASINO

The casino is the sharper analogy, because a casino is an environment engineered with an explicit behavioral objective and effectively no other.

The features are familiar and are worth restating in mechanism terms rather than in the language of vice. Clocks are absent, and windows are absent, so the passage of time cannot be estimated from the environment. The floor plan is deliberately non-obvious, so exit requires navigation and is never the default direction of travel. Money is converted into chips, so the unit being wagered is not the unit that hurts to lose. Losses are made quiet and wins are made loud, so the audible base rate of the room is far more favourable than the actual one. Reinforcement arrives on an unpredictable schedule rather than a reliable one.

Each of these has a direct counterpart in the taxonomy, and the correspondence is not loose. Temporal disorientation is **A-13**. Exit obstruction is **H-01** and **F-09**. Currency obfuscation, the chip, is **F-17**. Reinforcement on an unpredictable schedule is **C-02**, whose primary literature is Ferster and Skinner (1957).

A hard honesty note is required here, because the casino analogy is where persuasive-design writing most often over-claims. Ferster and Skinner established schedules of reinforcement in the operant laboratory, with animal subjects, under controlled conditions. The claim routinely made in product circles, that variable reward produces slot-machine-like compulsion in smartphone applications, is an *extrapolation* from that literature and not a finding within it. This document labels it as an analogy and does not attach any response-rate or resistance-to-extinction number to the Ferster and Skinner citation, because no such number has been retrieved from the text. The structural resemblance between a pull-to-refresh gesture and a lever pull is real and is worth stating. The inference that it produces the same behavior in humans at the same magnitudes is a hypothesis, and this document treats it as one.

#### WHAT TRANSFERS, AND WHAT IS WORSE

Three properties make a software environment a more powerful behavioral environment than either a supermarket or a casino.

**It is personalised.** A supermarket has one layout for every shopper. A feed has one layout per person, computed for that person, updated continuously. The gradient is not merely set; it is fitted.

**It is portable.** A casino has to wait for you to enter it. An application is in your pocket, and it can initiate the encounter. The notification is the moment the environment stops being a place you go and becomes a place that comes to you.

**It is instrumented.** A supermarket learns about its layout from till receipts, slowly and in aggregate. A software environment learns from every scroll, every hover, every abandoned session, per user, immediately. What is measured can be optimised, and what is optimised converges.

The environment argument, stated completely: **software is a choice architecture that is personalised, portable, and instrumented, whose gradients are set by a party with a measured interest in the outcome, and which therefore acts on behavior continuously rather than at the moment of use.**

---

### 1.4 Defining behavioral design

The literature does not offer a definition of the object this document studies. It offers definitions of a *subset* of it. Brignull's deceptive patterns, Gray et al.'s five strategies, Mathur et al.'s fifteen types, and the sixty-four-type ontology of Gray et al. (2024) are all definitions of the harmful case. There is no accepted term that covers the harmful case and the benign case together, and the absence of that term is a real problem: it means the field has no vocabulary in which to say that a streak in a language app and a streak that manufactures guilt to prevent departure are *the same mechanism* differently implemented.

The following definition is therefore this document's own.

#### Definition (author's own; proposed here).

**Behavioral design** *is the deliberate structuring of an interactive environment such that a target behavior becomes more probable than it would be under a minimal-intervention arrangement of the same functionality, where the increase in probability is achieved by acting on the user's attention, motivation, ability, or cost of alternatives rather than by changing the underlying value of the offer.*

Each clause is doing specific work, and the definition fails without any one of them.

**"Deliberate structuring."** Accidental effects are excluded from the definition, but not from responsibility. A designer who did not intend an effect has still produced it. What the clause distinguishes is the object of study: this document is about designs that were chosen. The fact that an unchosen design also has behavioral consequences is precisely why section 1.5 argues that there is no neutral interface.

**"Interactive environment."** Not "interface." The environment includes the notification that arrives when the interface is closed, the email sent after churn, the default that persists between sessions, and the state the product maintains about the user. Restricting the object to what is on screen would exclude most of Families A, C, and H.

**"Target behavior."** There is always one, and it is always nameable. Sign up. Return tomorrow. Grant the permission. Do not cancel. Invite a friend. A design without a target behavior is not behavioral design; it is decoration. Making the target behavior explicit is the first step of every teardown in this document, and the ability to state it is the test of whether a pattern has actually been understood.

**"More probable than under a minimal-intervention arrangement."** This is the counterfactual clause, and it is the hardest one. Behavioral design is defined relative to a baseline: what would this user have done if the same functionality had been presented in the least structured way that still works? Luguri and Strahilevitz's control condition is an unusually clean instance of exactly this baseline. Accept or decline, on one screen, symmetrically presented: 11.3 percent. Everything above that number is the measured contribution of the environment. Most of the time this baseline is unavailable to an outside analyst, which is a genuine methodological limitation of this document and is recorded as such in the Open Questions.

**"Acting on attention, motivation, ability, or cost of alternatives."** These four are the levers, and they exhaust the mechanism space at this level of abstraction. The middle two are Fogg's. Fogg (2009) proposes that behavior is a product of three factors, motivation, ability, and triggers (his term in 2009; he renamed triggers to prompts in late 2017), each with subcomponents. Attention is added here because Fogg's model presumes the user is present, and in consumer software the contest for presence *is* the first contest. Cost of alternatives is added because Fogg's model describes how to make a target behavior easier and says little about making the *competing* behavior harder, which is the entire mechanism of Families G and H. Sludge, obstruction, and exit friction operate on the alternative, not on the target.

A note demanded by the replication clause: **Fogg's 2009 paper contains no equation.** The formula "B = MAT" that circulates in product writing does not appear in it. The paper was retrieved and read in full; the model is stated in prose and in two figures only. The multiplicative form implies a mathematical claim Fogg never made. Anywhere this document uses Fogg, it uses the prose formulation, and any framework of ours that resembles his is *inspired by* him and says so.

**"Rather than by changing the underlying value of the offer."** This is the exclusion clause, and it is what stops the definition from swallowing all of product management. A team that makes a product genuinely faster, cheaper, or more useful has changed behavior, but it has done so by changing the offer. That is competition, and it is not the subject of this document. Behavioral design is what moves behavior with the offer held constant. This is exactly the axis Luguri and Strahilevitz isolated. The service did not improve between the 11.3 percent condition and the 41.9 percent condition. Only the environment did.

Two things the definition deliberately does **not** do.

It does not mention deception. A pattern can be entirely honest and still be behavioral design. Endowed progress (**B-03**) tells no lie.

It does not mention harm. Whether a given instance of behavioral design harms the user is a separate question, answered on a separate axis, and section 1.10 is where that axis is introduced. Building harm into the definition is the mistake that produced four taxonomies which cannot describe the majority of what designers actually do.

---

### 1.5 Every interface changes behavior

The strong form of the argument in this chapter is that there is no such thing as a neutral interface, and this section defends that claim rather than assuming it.

The defence is not that all designers are manipulators. It is arithmetic. Any interface must make a finite set of choices, an order, a default, a size, a wording, a number of steps, and every one of those choices has a measurable behavioral consequence, and there is no setting of them that has no consequence. The neutral interface would be one in which changing any of these variables changed nothing about what users do. No such interface exists, because if it did, no A/B test would ever return a result.

The following are interface-to-behavior chains. Each states a concrete design decision, the behavior it produces, and the mechanism by which it does so. Where the mechanism rests on a cited effect, the citation is given. Where it does not, it is marked.

**Chain 1: default → adoption.** An option is preselected. More users end up with it than would have chosen it. The mechanism is that inaction is cheaper than action, and a default converts adoption into inaction. This is **G-02** in the taxonomy, and it is the most efficient behavioral intervention available to a designer, because it requires the user to do nothing at all. Luguri and Strahilevitz's Study 2 provides the closest verified support: among the strategies tested, hidden information, trick questions, and obstruction were the most effective at manipulating consumers, and all three operate by making the user-preferred path more expensive than the product-preferred one. That precise ordering is confirmed; the per-strategy percentages are not, and are not printed here.

**Chain 2: button asymmetry → consent.** Accept is a filled, coloured, large button. Decline is grey text of lower contrast, or is one level down in a submenu. Acceptance rises. The mechanism is that the two options are not being compared on their merits; they are being compared on their salience and their cost in clicks. This is **G-01** and **G-08**. Its ethical counterpart, **I-02** (Symmetric Consent), is defined as giving both options equal visual weight and equal cost, which is the same mechanism with the gradient removed.

**Chain 3: removing the page boundary → session length.** A paginated feed ends. An infinite feed does not. Session length increases. The mechanism is that a page boundary is a *stopping cue*, a moment at which continuing requires a positive decision. Remove it, and continuing becomes the default while stopping becomes the action requiring effort. This is the inversion that defines **A-01**. Note that no deception is involved at any point, and no user is prevented from stopping. The design has not changed what is possible. It has changed which behavior is the one that costs something. Empirical magnitude of the session-length increase: CITATION NEEDED.

**Chain 4: reference price → willingness to pay.** A higher price is displayed next to the price actually being offered. The offered price is judged as smaller than it would be in isolation. The mechanism is evaluation against a reference point rather than in absolute terms, which is the central structural claim of prospect theory: value is defined on gains and losses relative to a reference point rather than on final states (Kahneman & Tversky, 1979). This is **F-05**.

Prospect theory deserves particular emphasis here, because it is the sturdiest item in this chapter's evidence base. Its patterns were tested for replication by Ruggeri et al. (2020) across 4,098 participants in 19 countries and 13 languages; the results replicated for 94 percent of items, with some attenuation, and twelve of thirteen theoretical contrasts replicated. In a document that is required by its own constitution to report replication status, it is worth being explicit: this is the one place in this chapter where the underlying science can be leaned on hard.

**Chain 5: framing as loss → action.** "You will lose your streak" outperforms "Continue your streak." The mechanism is the asymmetry of the value function, which Kahneman and Tversky (1979) state as being "generally steeper for losses than for gains." A prospective loss is weighted more heavily than an equivalent prospective gain. Note carefully what is and is not being claimed. The value-function asymmetry is verified from the 1979 primary text and replicates. The *extension* of it to streak-loss

messaging in a mobile application is this document's inference, not a cited finding, and is labelled accordingly.

**Chain 6: unfinished state → return.** A profile is 60 percent complete and says so. Users come back and complete it. The mechanism is the pull of the interrupted task. And here this document must break with almost the entire UX literature, because that literature attributes this to the Zeigarnik effect, and the Zeigarnik effect does not replicate. Ghibellini and Meier's 2025 meta-analysis found the pooled ratio of recall for interrupted versus completed tasks to be 0.99, effectively no memory advantage at all, and concluded that "the Zeigarnik effect lacks universal validity." What *does* replicate is the **Ovsiankina effect**: the tendency to *resume* an interrupted task, with a pooled resumption rate of 67.0 percent. The correct mechanism for a progress bar was never memory. It was resumption. Every citation of Zeigarnik in support of progress indicators is citing the wrong effect, and it is the neighbouring effect, the one that survives meta-analysis, that actually explains the behavior. This is **C-13**.

**Chain 7: granted progress → persistence.** A loyalty card requiring eight steps is reissued as a card requiring ten steps with two already completed. The task is identical in the work it demands. Completion rates rise and completion times fall. Nunes and Drèze (2006) document exactly this manipulation and name the phenomenon the endowed progress effect. The mechanism, stated explicitly in their abstract and routinely misreported in product writing, is *perceived task completion*, not sunk-cost avoidance. This is **B-03**. The widely circulated completion percentages from this study are not confirmed against the primary text and are not printed here.

Seven chains, seven mechanisms, and not one of them requires a lie. That is the finding of this section. **Deception is one technique within behavioral design. It is not the definition of it, and it is not even the majority of it.** A taxonomy organised around deception can describe Chain 2 and struggles to describe Chains 3, 4, 6, and 7 at all, which is why this document is organised around mechanism instead.

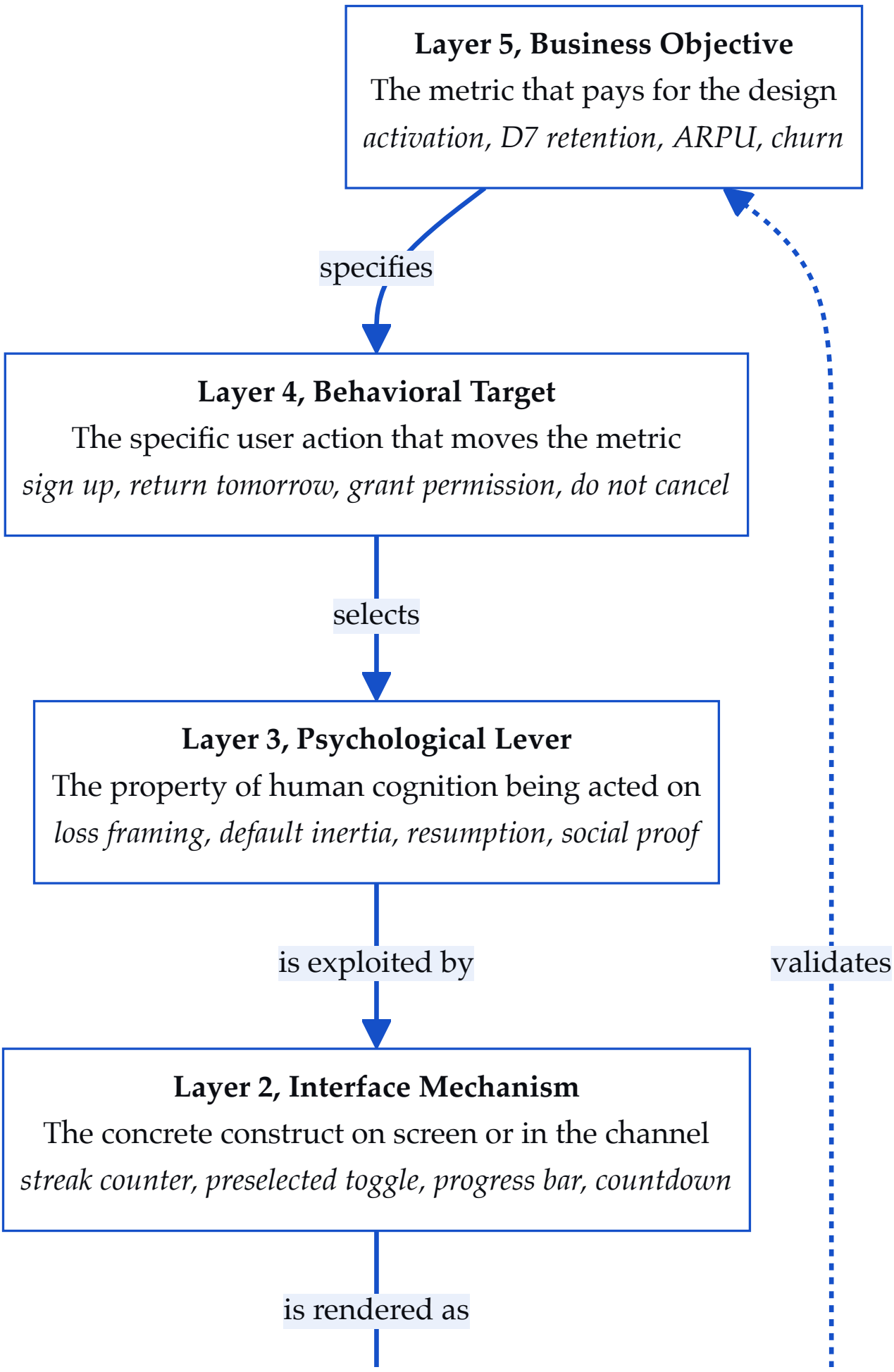
---

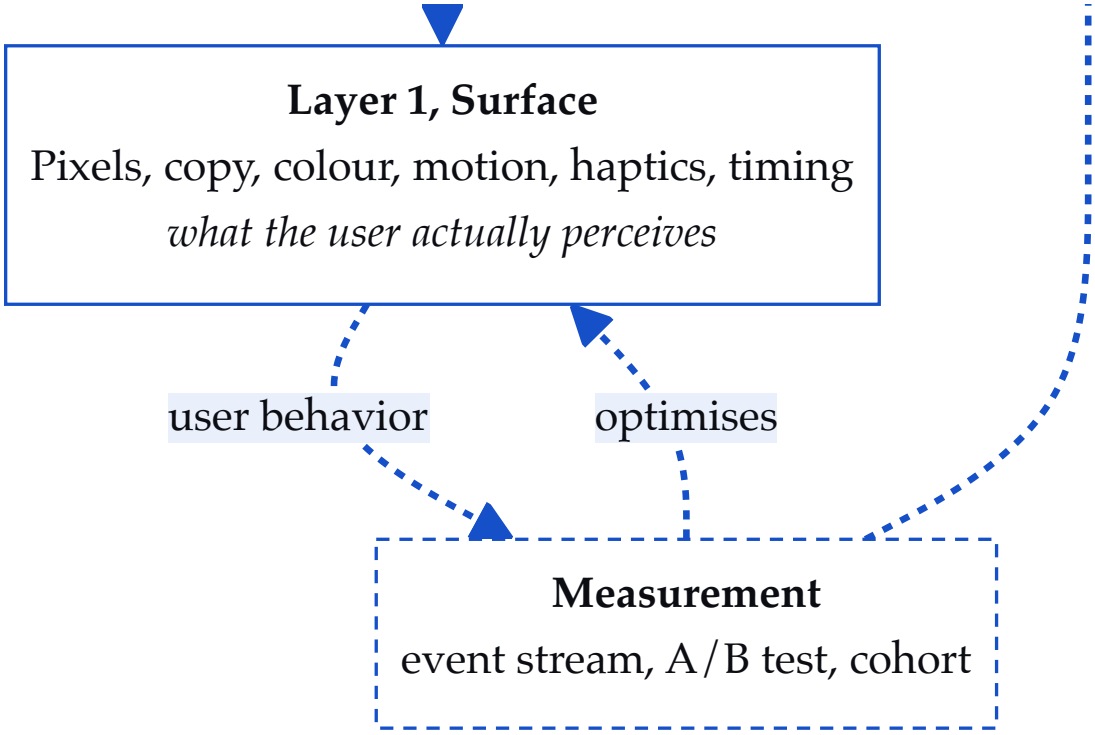
### 1.6 The Behavioral Design Stack

*Original framework; proposed here.*

The interface-to-behavior chains above have a shared structure that is worth extracting. Every one of them runs from something the company wants, through something a designer built, through something in the human mind, to something the user did, and back to something the company counted. The Behavioral Design Stack is the name this document gives to that structure.

The stack has five layers. Its purpose is diagnostic: given any pattern, one should be able to name what sits at each layer, and a pattern for which a layer cannot be named has not been understood.


Layer 5, Business Objective
The metric that pays for the design
activation, D7 retention, ARPU, churn
specifies
Layer 4, Behavioral Target
The specific user action that moves the metric
sign up, return tomorrow, grant permission, do not cancel
selects
Layer 3, Psychological Lever
The property of human cognition being acted on
loss framing, default inertia, resumption, social proof
is exploited by
Layer 2, Interface Mechanism
The concrete construct on screen or in the channel
streak counter, preselected toggle, progress bar, countdown
is rendered as
validates




Three claims about this structure.

**The stack is built top-down and experienced bottom-up.** The company begins at Layer 5 with a number it must move and works downward to a surface. The user begins at Layer 1 with a surface and never sees the rest. This asymmetry of visibility is not incidental; it is a precondition for the manipulative valence defined in section 1.10, which is specifically the case that *works because the user cannot see the upper layers*.

**The dashed feedback loop is what makes software different.** A supermarket has Layers 1 through 5. It does not have the loop, or it has a slow and coarse version of it. The loop is the mechanism by which a design that was merely *hypothesised* to move a metric becomes a design that is *known* to move it, and then becomes a design that is refined until it moves it as much as possible. Every pattern in the taxonomy that has a manipulative valence got there by being optimised, not by being invented that way. This is the process Luguri and Strahilevitz point to when they write that dark patterns proliferate "because firms' proprietary A-B testing has revealed them to be profit maximizing."

**Ethical analysis at the wrong layer fails.** Critiques of behavioral design habitually operate at Layer 2, producing lists of forbidden mechanisms: no countdowns, no streaks, no infinite scroll. This cannot work, because the same Layer 2 mechanism serves ethical and manipulative Layer 5 objectives without changing its appearance at Layer 1. The pressure that produces a harmful pattern originates at Layer 5. A regulation, or a design principle, that addresses only Layer 2 will be routed around by a team whose Layer 5 objective is unchanged, and it will be routed around within a quarter, because the feedback loop makes finding the substitute cheap. This is the central structural argument for why this document is organised by mechanism and valence rather than by prohibition, and it is developed in full in section 1.11.

## 1.7 Three levels of design

Design as a discipline has developed powerful, teachable, well-instrumented theory at two levels and almost none at a third. The gap is the space this document is written into.

#### LEVEL 1: VISUAL DESIGN

The question: *does it look right?*

The unit of analysis is the composition. Typography, hierarchy, colour, spacing, contrast, alignment. This level is mature. It has centuries of prior art in print, an inherited grammar, formal training pipelines, style guides, design systems, and, importantly, automated checkers. Contrast ratios can be linted. Token drift can be detected in continuous integration. A junior designer can be told, with precision and without appeal to taste, that a thing is wrong.

#### LEVEL 2: INTERACTION DESIGN

The question: *does it work?*

The unit of analysis is the task. Can the user complete what they came to do, quickly, with few errors, and understand what happened? This level is also mature. It has heuristics, usability testing protocols, task-success and time-on-task metrics, error-rate instrumentation, accessibility standards with legal force. A team that ships a confusing flow will find out.

#### LEVEL 3: BEHAVIORAL DESIGN

The question: *what does it make people do, over time, and in whose interest?*

The unit of analysis is not the composition or the task. It is the **behavior**, and its timescale is not the session. It is weeks, months, and years. And at this level the discipline has almost nothing.

There is no behavioral equivalent of a contrast ratio. There is no lint rule that fires when a design creates a compulsion. There is no accepted metric for the cost a pattern imposes on a user, no standard for how much friction on an exit path is too much, no professional certification that covers it, and, with the recent and partial exception of consent interfaces, no regulatory standard with teeth. A designer can ship an interface that is flawless at Level 1, exemplary at Level 2, and predatory at Level 3, and every review process the organisation possesses will pass it.

This is not a rhetorical point. It is a claim about the actual state of the field, and it can be evidenced from the structure of the literature itself.

Consider what the existing scholarship on behavioral design consists of. Brignull's deceptive.design lists 18 types. Gray et al. (2018) analysed a corpus of 118 practitioner- identified artifacts and derived five *strategies*: nagging, obstruction, sneaking, interface interference, and forced action. Mathur et al. (2019) crawled approximately 11,000 shopping websites and roughly 53,000 product pages, found 1,818 dark-pattern instances across 15 types in 7 categories on 183 deceptive sites, and identified 22 third-party entities selling dark patterns as a turnkey service. Gray et al. (2024) harmonised ten existing taxonomies into a three-level ontology of 64 types.

Every one of these is excellent work, and this document depends on all of them. But note what they all are: **catalogues of harm**. They enumerate what should not be done. Not one of them is a theory of how

interfaces produce behavior in general, and none of them has anywhere to put a design that helps the user, because a catalogue of deception has no cell for a non-deception. Level 3 has a growing pathology and no physiology.

The three levels also differ in a way that is not usually stated and that matters for professional practice.

| | LEVEL 1, VISUAL | LEVEL 2, INTERACTION | LEVEL 3, BEHAVIORAL |
|---|---|---|---|
| Question | Does it look right? | Does it work? | What does it make people do? |
| Unit | Composition | Task | Behavior |
| Timescale | Instant | Session | Months to years |
| Failure is | Visible | Reported | Invisible |
| Who notices | Anyone | The user | Often no one |
| Standards | Mature | Mature | Effectively absent |
| Automated checks | Yes | Partial | None |

The row that carries the argument is **"failure is invisible."** An ugly interface is obvious to everyone. A confusing interface produces support tickets. A behaviorally harmful interface produces *engagement*, which the organisation's instruments are configured to read as success. The feedback signal that would tell the team something is wrong is precisely the signal that tells them things are going well. A discipline in which the failure mode is indistinguishable from the success metric will not self-correct, and has not.

---

### 1.8 The Behavior Loop

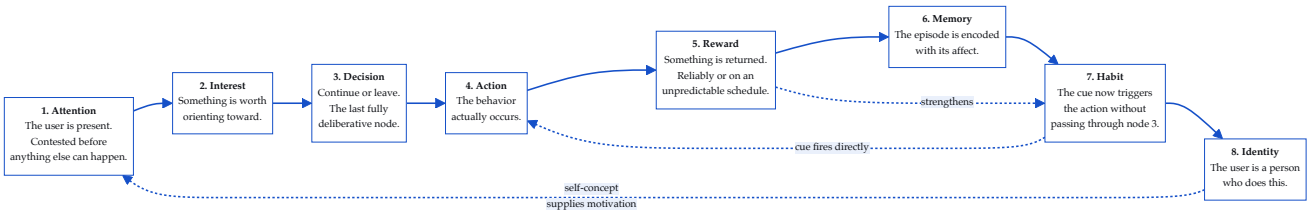
*Original framework; proposed here. Inspired by, but not derived from, Fogg (2009); it is not the Hook Model.*

Sections 1.5 through 1.7 established that interfaces change behavior and that the discipline lacks a theory of how. This section proposes one, at the level of the individual user, over time.

Two disclaimers first, both required.

This loop is **not** Fogg's Behavior Model. Fogg (2009) proposes that behavior is a product of motivation, ability, and triggers, and his model explains a *single* behavior at a *single* moment. That is a different object from the one modelled here, which is the progression of a behavior across repetitions. Fogg's paper contains no equation and this document never writes one on his behalf.

This loop is also **not** Eyal's Hook Model. The Hook Model, trigger, action, variable reward, investment, is a practitioner framework with no primary empirical base of its own, and it terminates at investment. The loop below is longer at both ends, and the two nodes it adds at the far end are the ones that matter most for the ethical argument in this document.




### NODE BY NODE

**1. Attention.** Nothing in the loop can begin until the user is present, and presence is scarce. This is why Family A exists as a separate family and why it is placed first in the taxonomy: capture precedes persuasion. Fogg's model presumes the user is already at the device. Consumer software cannot presume that, and the notification is the industry's answer to the problem. Patterns: **A-01** through **A-13**.

**2. Interest.** Attention is oriented but not committed. The user is scanning. This node is where relevance, novelty, and social signal do their work, and where the algorithmic feed (**A-09**) has its leverage: it does not need to persuade the user of anything, only to place in front of them the item with the highest predicted probability of holding them.

**3. Decision.** The user chooses. This is the last node in the loop at which the process is fully deliberative, and consequently it is the node at which the largest concentration of behavioral machinery is aimed. Defaults act here (**G-02**). Framing acts here (**F-05**). Loss aversion acts here (Kahneman & Tversky, 1979). Sludge acts here, by making the alternative expensive rather than the target attractive. Everything from Luguri and Strahilevitz's experiments happened at this node, and the size of the effect they found, 11.3 percent to 41.9 percent, product and price held constant, is the best available measure of how much a single node of this loop can be moved by environment alone.

**4. Action.** The behavior occurs. Ability governs here, which is Fogg's contribution: his six simplicity factors are time, money, physical effort, brain cycles, social deviance, and non-routine. Reducing any of them raises the probability of the action.

**5. Reward.** Something is returned to the user. The schedule matters as much as the content. Ferster and Skinner (1957) established that reinforcement delivered on an unpredictable schedule produces more persistent behavior than reinforcement delivered reliably. The application of that finding to software is **C-02**, and, as stated in section 1.3 and repeated here because it matters, the transfer of the operant finding to human smartphone behavior is an analogy, not a demonstrated result.

**6. Memory.** The episode is encoded, with its affect attached. This is where a document that has not verified its sources would reach for the Zeigarnik effect and claim that incomplete episodes are better remembered. They are not: pooled recall ratio 0.99 (Ghibellini & Meier, 2025). What is encoded and matters is not superior recall of the incomplete but the *affective tag* on the episode, and, separately, and robustly, the disposition to **resume**, at a pooled rate of 67.0 percent. The unfinished thing does not haunt the user's memory. It pulls at their behavior. Those are different claims and only the second one survives meta-analysis.

**7. Habit.** The critical transition of the entire loop. A habit is the state in which the cue triggers the action **without passing through node 3**. The decision node is bypassed. This is the dashed edge from Habit back to Action, and it is drawn dashed because it is the edge on which the ethics of the whole system

turn: a habit is, by construction, a behavior that no longer routes through deliberation. Eyal, whose framework is a practitioner framework and is cited as such, quotes cognitive psychologists in defining habits as "automatic behaviors triggered by situational cues," and states the design goal plainly: to attach the product to an internal trigger so that "users show up without any external prompting." That is an accurate description of what habit-forming design is trying to achieve. Whether a designer *should* be trying to route a user's behavior around their own deliberation, and under what conditions, is the question that section 1.10 exists to answer. It is not answered by pretending the goal is something else.

**8. Identity.** The terminal node, and the one absent from every prior model this document is aware of. The user no longer merely *does* the behavior; they are a person *who does* it. "I'm a runner." "I have a 900-day streak." "I'm a Wordle person." At this node the product no longer has to supply motivation, because the self-concept supplies it, which is the dashed edge back to Attention.

Identity is included because it is where the loop's ethical stakes are highest in both directions. It is the node at which behavioral design becomes genuinely valuable to a user, since a person who has become the kind of person who exercises has gained something real, and a product that helped them get there has done something good. It is also the node at which leaving the product becomes a loss of self rather than a change of software, which is the mechanism underlying **D-07** (Identity Investment), **D-12** (Reputation Lock-in), and the guilt messaging of **C-14**. **The same node that makes behavioral design worth doing is the node that makes it dangerous.** No other framework in this space models it, and this is the principal reason the loop is proposed.

#### WHAT THE LOOP CLAIMS AND DOES NOT CLAIM

It claims that the eight nodes are traversed in order on a first encounter, that repetition strengthens the direct Habit-to-Action edge, and that the loop's ethical character is determined by whether the user, if shown the diagram, would endorse being moved along it.

It does not claim that the nodes are neurologically distinct, that the loop is complete, or that it has been empirically validated. It has not been. It is a framework for organising the taxonomy, offered as a contribution and testable in principle, and the design of an experiment that would test it is recorded in the Open Questions.

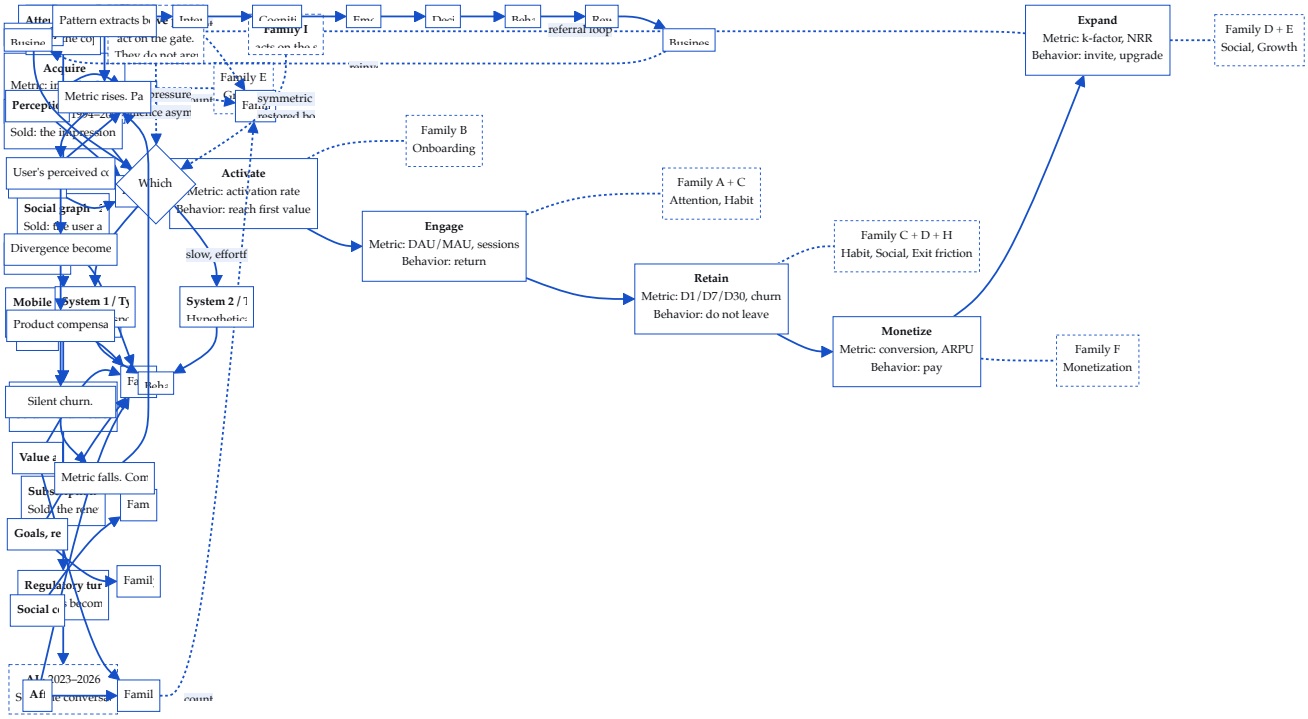
---

### 1.9 The consumer software engine

The Behavior Loop describes one user. It does not explain why any of it gets built, and a theory that cannot explain why a design exists cannot predict what will be designed next.

Behavior does not appear in a product because a designer found it interesting. It appears because it serves a business metric. The following is the engine those metrics sit in.

*Original framework; proposed here. The stage names are the common vocabulary of the consumer software industry; the mapping to behavioral loops and to the taxonomy's families is this document's own.*




The important reading of this diagram is the direction of causation, and it runs the opposite way from the way designers describe their own work.

A designer will say: *we built a streak because we wanted to help users form a daily habit.* The engine says: *a streak was built because D7 retention was the quarter's objective, and a streak is the cheapest known intervention that moves D7.* Both sentences can be true, and frequently both are. But only the second one predicts what will be built next, because only the second one names the thing that is actually being optimised.

**The behavior exists to serve the metric. It does not exist for its own sake, and it does not exist for the user's sake, even when it happens to serve the user.** This is not a cynical claim. It is the load-bearing claim of section 1.11, and it is the reason that appeals to designer conscience are a weak intervention: the conscience is at Layer 2 of the stack, and the pressure originates at Layer 5.

Three properties of the engine matter for everything that follows.

**Every stage has a corresponding family, and the map is nearly complete.** This is not a coincidence and it is not a construction artifact. The taxonomy's nine families were derived from mechanism, and the engine's six stages from business practice, and they align because each stage generates a behavioral demand and the families are the supply. The one family with no natural home in the engine is Family I, Trust and Autonomy Preservation, and the reason is diagnostic: **the engine has no stage whose metric is served by the user retaining autonomy.** Family I exists in the taxonomy and has no cell in the engine. That absence is a finding, and it is the sharpest single statement of the problem this document is about.

**The Retain stage is where the pressure concentrates.** Acquisition costs money. Retention protects money already spent. A churning user is not a neutral event; it is a realised loss on an acquisition already paid for, and Kahneman and Tversky's value function tells us that a loss is weighted more

heavily than an equivalent gain by the *person* evaluating it. Whether the same asymmetry applies to an organisation evaluating a cohort is an inference this document makes and cannot cite; it is marked. CITATION NEEDED. What is not an inference is the structural observation: Families H (Exit Friction), C (Habit), and D (Social) between them contain 38 of the taxonomy's 116 patterns, and the highest severity ceilings outside monetization cluster there. The stage with the most to lose has the most machinery.

**Monetization is where the mechanisms become adversarial in the clearest sense.** At the Engage stage, the company wants the user to return, and the user, at least sometimes, also wants to return. Interests partially align. At the Monetize stage, every marginal rupee or dollar extracted from the user is a rupee or dollar the user does not keep. The zero-sum component is larger, and the taxonomy reflects this: Family F is the largest family, with 18 patterns, and it contains the greatest concentration of severity-5 ceilings. Luguri and Strahilevitz's experiments were conducted at exactly this stage, a subscription acceptance decision, and found the environment capable of moving acceptance from 11.3 percent to 41.9 percent with the offer held constant. If the environment can do that where the interests are most opposed, the mechanism is not a marginal one.

---

### 1.10 The ethical spectrum, and the concept of valence

This is the spine of the document. Everything before it establishes that software changes behavior. Everything after it asks whether a specific change is acceptable. This section is the joint.

The dominant framing in both public discourse and the academic literature is binary: there are dark patterns, and there is everything else. The literature is explicit about this. Gray et al. (2018) define dark patterns as instances "where user value is supplanted in favor of shareholder value", a definition that is exactly correct for the harmful case and has no capacity to describe anything else. Every taxonomy in the field inherits this binary structure and is therefore, by construction, a list of prohibitions.

The binary fails for a reason that is easy to state and hard to design around: **the same mechanism is ethical in one implementation and manipulative in another, without changing appearance.**

A streak (**C-01**) is a count of consecutive periods of use. In a language-learning application, used by a person who has decided they want to study daily and who is struggling to do so, the streak supplies exactly the external structure the user has consciously asked for. It serves a goal the user endorses. Delete the streak and the user is worse off by their own account.

The same counter, in a social application, generating escalating messages that frame a lapse as a failure toward one's friends, is **C-14**. It is not helping the user reach a goal the user set. It is manufacturing an emotion in order to prevent an outcome the *company* does not want. The user, if asked, would not endorse being kept this way.

A taxonomy that files "streak" permanently under "dark patterns" cannot express this difference, and therefore cannot describe how real products actually work. So this document puts mechanism and morality on different axes, and gives the moral axis a name.

VALENCE

**Valence** *is a property of an implementation, not of a mechanism. It has three values.*

**Ethical.** *The user's interest and the business interest align, and the user retains autonomy. The design would survive full disclosure. If the user were shown the Behavioral Design Stack for this pattern, including Layer 5 and the metric it serves, they would endorse it anyway.*

**Persuasive.** *The business interest leads. The user's autonomy is preserved and disclosure is honest. The design is trying to get the user to do something the business wants, it is not hiding that fact, and the user retains a real and inexpensive ability to decline. Advertising, at its most honest, lives here. So does most of Family B.*

**Manipulative.** *The design works because the user does not understand what is happening to them, or would object if they did. Concealment is not incidental to the mechanism; it is load-bearing. Disclosure would defeat it.*

*(Definitions per PROJECT.md, section 4. The valence axis is this document's original contribution.)*

The test that separates Persuasive from Manipulative is the one worth internalising: **would disclosure defeat it?**

A price anchor still works when the user knows what an anchor is. It is persuasive. A countdown timer that silently resets does not work on a user who has been told it resets; its entire efficacy depends on the user believing it is real. It is manipulative. This is not a difference of degree along a single scale of unpleasantness. It is a categorical difference in the mechanism's relationship to the user's understanding, and it is why manipulation and persuasion require different remedies. Persuasion is answered by competition. Manipulation is answered by disclosure or by law, because the user cannot answer it themselves, by construction, they cannot see it.

#### WHY VALENCE IS A PROPERTY OF THE IMPLEMENTATION

This point is easy to state and constantly violated in practice, so it is worth stating formally.

The mechanism is: *a visible count of consecutive days of use, whose value derives from its fragility.* That is a fact about an interface. It has no ethics.

The implementation is: *this count, in this product, serving this metric, with this exit cost, addressed to this user who has or has not asked for it.* That is where the ethics live, and every term in it can vary without the mechanism changing at all.

Three consequences follow, and each is structural rather than rhetorical.

**Valence cannot be determined from a screenshot.** It requires knowing what the pattern is for, what happens if the user declines, and whether the user would endorse it on full disclosure. This is a real methodological burden on this document, it means every pattern entry must argue its valence rather than assert it, which is why the pattern template in PROJECT.md has a mandatory "Valence analysis" section, and it is a real limitation on automated detection. Mathur et al. (2019) were able to detect 1,818 dark-pattern instances at scale across roughly 11,000 sites precisely because they restricted

themselves to types whose harmfulness is inferable from the artifact alone. That restriction is what made the crawl possible, and it is also the boundary of what a crawl can ever find.

**A pattern's valence is not fixed over its lifetime.** It is set by a business objective, and business objectives change. A design that was ethical when the company was growing can become manipulative when the company is defending revenue, without a single line of code changing. The pattern did not move. Layer 5 moved beneath it. This is a prediction the framework makes and, as far as this document is aware, an untested one: it is recorded in the Open Questions.

**There is a family of patterns with no manipulative implementation, and existing taxonomies have nowhere to put it.** Family I of the taxonomy, Trust and Autonomy Preservation, is twelve patterns that achieve legitimate business goals without extracting from the user: symmetric consent (**I-02**), frictionless exit (**I-03**), the stopping cue that restores the boundary infinite scroll removed (**I-06**), the unused-value alert that tells a user they are paying for something they do not use (**I-10**). No catalogue of deception can contain these, because a catalogue of harms has no cell for a design that helps.

This is where the document's normative argument is actually made, and it is worth being explicit that it is *not* made by condemnation. It is made by demonstrating that for each harmful pattern, the legitimate business goal it serves has a non-harmful solution. A company that ships **H-01** (Cancellation Maze) has not merely done something wrong. It has chosen the harmful solution to a problem that has a non-harmful one, and Family I is the document's attempt to prove that the non-harmful one exists.

The honest counterpoint, recorded here and in the Open Questions because it is the hardest evidentiary burden in the project: Family I risks being a wish-list. Each entry needs a shipped, named product that actually does it, or it should be cut.

---

### 1.11 Why patterns exist

The most common explanation for manipulative design is that the people who build it are bad people. This explanation is emotionally satisfying, occasionally true, and analytically useless, because it predicts nothing. If harmful patterns were produced by bad actors, we would expect them to be concentrated in disreputable companies. They are not. Mathur et al. (2019) found dark patterns distributed across a crawl of roughly 11,000 shopping websites, including mainstream ones, and, the more telling finding, identified **22 third-party entities offering dark patterns as a turnkey solution**. Manipulative design is not a deviance. It is an *industry*, with vendors, integration documentation, and presumably customer support.

The better explanation is that manipulative design is the equilibrium output of a system whose incentives are correctly aligned to produce it. Consider the pressures acting on a product team that contains no bad actors at all.

**Pressure 1: the metric is proximate and the harm is distal.** The team is accountable for D7 retention this quarter. The harm from a guilt-inducing streak notification is diffuse, delayed, distributed across millions of users, and lands on none of the team's instruments. A cost that is not measured is,

organisationally, a cost that does not exist. It is not that the team weighed the harm and accepted it. It is that the harm was never on the scale.

**Pressure 2: the loop optimises toward the gradient, not toward a decision.** Return to the Behavioral Design Stack and its dashed feedback edge. Nobody at the company convened a meeting and resolved to build a manipulative pattern. Somebody ran a test in which the decline button was grey instead of blue. It won. It shipped. Somebody else, later, tested moving the decline link into a submenu. It won. It shipped. Each individual step is a one-percent improvement that any reasonable person would approve, and the composition of several hundred such steps is a consent flow that no reasonable person would design from scratch. **Manipulation is reached by gradient descent, and it is reached by teams that never decided to go there.** This is why exhorting designers to have better values is a weak intervention: the mechanism does not run through anybody's values. It runs through an optimiser.

**Pressure 3: it works, and the evidence is proprietary.** Luguri and Strahilevitz put the mechanism precisely: dark patterns are "presumably proliferating because firms' proprietary A-B testing has revealed them to be profit maximizing." A team that declines to use them is not making a costless ethical choice. It is choosing, measurably, to convert 11.3 percent where a competitor converts 41.9 percent. Under competition, that is not a stable position. This is a **race to the bottom in a market for behavior**, and the individual firm's incentive to defect is exactly as strong as it is in any other such race.

**Pressure 4: nobody in the loop represents the cost.** The company has a growth team representing the metric, a legal team representing regulatory risk, and a design team representing usability. There is no function whose job is to represent the user's time, attention, or autonomy, none of which appear on any dashboard. The absence is not a conspiracy. It is an org chart.

**Pressure 5: the user cannot punish what they cannot see.** In a functioning market, a bad product is punished by exit. But the manipulative valence is *defined* by the fact that the user does not perceive the mechanism. A user who has been steered by a default has, by construction, not noticed being steered, and therefore does not attribute the outcome to the company, and therefore does not leave. The market's corrective mechanism requires perception, and manipulation is precisely the class of harms that operates below it. **This is the argument for why this cannot be left to the market, and it is a structural argument, not a moral one.**

The conclusion is uncomfortable and is the reason this document exists in the form it does:

### Harmful behavioral design is not primarily a failure of ethics. It is the correct output of a system that measures engagement, does not measure harm, iterates faster than users can adapt, and operates in a market where restraint is unilaterally costly.

If that is true, then interventions aimed at individual designer virtue will fail, because they act on Layer 2 of a stack whose pressure originates at Layer 5. So will lists of forbidden patterns, because the optimiser will route around any specific prohibition within a quarter, the loop makes finding a substitute cheap, and because a prohibition list cannot distinguish the ethical implementation of a mechanism from the manipulative one and will therefore either ban too much or be ignored.

What might work has to change what is measured, what is disclosed, or what is priced. That is a claim this chapter asserts and does not yet defend. The defence, and the counter-argument that it is naive,

are Chapter 9.

Two things stated for the sake of honesty. First: the argument above is *exculpatory of individuals and not of firms*. That an outcome is the equilibrium of a system does not absolve the participants who construct and profit from the system, and it does not make the harm smaller. Second: the harm is not hypothetical. Regulators have begun to price it. Gray et al. (2024) record the $245 million USD FTC judgment against Epic Games as an enforcement example, and note that the term "dark patterns" was codified into EU law in 2022 via the Digital Services Act, the Digital Markets Act, and the Data Act proposal, and into US law in the California CPRA; India's Department of Consumer Affairs released draft dark-pattern guidelines in the summer of 2023 and finalised them in November 2023.

---

## 1.12 Research objectives and contribution

THE QUESTION

This document answers one question:

### How does consumer software change human behavior, and by what mechanisms?

Chapter 1 has answered the first half. Software changes behavior because it is not a tool but an environment; because environments have gradients; because those gradients are set by a counterparty with a measured interest in the outcome; and because, uniquely among environments, this one is personalised, portable, instrumented, and optimised continuously against a behavioral objective. The magnitude of the effect is not speculative: with product and price held constant, the arrangement of the environment alone moved acceptance of a service from 11.3 percent to 41.9 percent in a representative sample of American consumers (Luguri & Strahilevitz, 2021).

The second half, *by what mechanisms*, is the taxonomy, and it is the rest of the document.

WHAT THE LITERATURE HAS, AND WHAT IT LACKS

| WORK | TYPES | ORGANIZING QUESTION | COVERS NON-HARMFUL PATTERNS? |
|---|---|---|---|
| Brignull, deceptive.design (2010–) | 18 | How is the user deceived? | No |
| Gray et al., CHI 2018 | 5 strategies | What is the designer's strategy? | No |
| Mathur et al., CSCW 2019 | 15 types / 7 categories | What is empirically observable at scale? | No |
| Gray et al., CHI 2024 ontology | 64 types, 3 levels | How do we harmonise 10 taxonomies? | No |
| **This work** | **116, 9 families, 3 axes** | **What is the mechanism, and whose interest does this implementation serve?** | **Yes, Family I** |

Every figure in that table is verified against a retrieved source.

The right-hand column is the gap. Four taxonomies, all of them useful, all of them catalogues of deception, and therefore all of them structurally unable to describe the ethical and persuasive majority of what behavioral designers actually do. A regulator needs a list of prohibitions. **A designer needs a theory**, and cannot get one from a list of things not to do.

A caution the honesty of this document requires: **116 and 64 are not comparable numbers and must never be presented as though the larger one is better.** Our count is larger not because we found more ways to harm users but because we admit patterns that are not harms. Gray et al. (2024) harmonised ten taxonomies to converge on a shared regulatory vocabulary. That is a different and in several respects harder job than ours, and where our names overlap theirs, we defer to theirs.

### OBJECTIVES

1. **Describe the mechanism by which interactive environments change behavior**, at a level

of abstraction general enough to cover the benign and the harmful case with one theory. This is Chapter 1.

1. **Build a taxonomy organised by mechanism rather than by harm**, with harm as an

independent axis. 116 patterns, 9 families, 3 axes.

1. **Separate mechanism from morality** by making valence a property of the implementation.

This is the load-bearing structural decision of the work, and everything else follows from it.

1. **Demonstrate that the harmful solution is never the only solution**, by giving every

harmful pattern an ethical counterpart in Family I that serves the same business goal.

1. **Ground every empirical claim, or mark it.** A short honest bibliography beats a long

fictional one, and this chapter's CITATION NEEDED markers are left visible on purpose.

1. **State replication status as part of the citation.** Behavioral science has a replication

crisis and this document is about behavioral science. Where the effect everyone cites does not replicate, Zeigarnik, we say so, and cite the one that does.

### CONTRIBUTION

Six things this document offers that, to the best of the author's knowledge, the existing literature does not.

**1. The valence axis.** Mechanism and morality separated onto different axes, so that the same pattern appears once rather than twice and its ethical status becomes a property of the implementation. No existing taxonomy does this, and none of them can express the difference between **C-01** and **C-14**.

**2. Family I.** Twelve patterns that are *good*. No dark-pattern taxonomy has a family of these, because a catalogue of harms has nowhere to put a design that helps. This is where the normative argument is

made, constructively, by demonstrating that the non-harmful solution exists, rather than by condemnation.

**3. The Behavioral Design Stack** (section 1.6). A five-layer diagnostic that locates where in an organisation a pattern originates, and thereby explains why interventions aimed at the wrong layer fail.

**4. The Behavior Loop** (section 1.8). An eight-node model whose terminal node, Identity, is absent from every prior model the author is aware of, and which is where the ethical stakes of behavioral design are highest in both directions.

**5. The economic argument** (section 1.11). Harmful patterns explained as the equilibrium output of a correctly functioning optimisation system rather than as a failure of individual ethics, with the corollary that designer-virtue interventions and prohibition lists will both fail, and for the same reason.

**6. Replication honesty as method.** A report that says "the effect everyone cites here does not replicate, and here is the one that does" is stronger than one that repeats the folklore. Two corrections in this chapter, Fogg's non-existent equation, and Zeigarnik's non-replicating memory effect, are small original contributions in themselves, because both errors are near-universal in the practitioner literature this document is written against.

---

### Key Takeaways

1. **Software is an environment, not a tool.** It has behavioral gradients, they were set by

someone, and that someone has a measured interest in the outcome.

1. **Decision architecture can outweigh price.** Holding product and price constant and

varying only the interface moved acceptance of a service from 11.3 percent (control) to 25.8 percent (mild dark patterns) to 41.9 percent (aggressive dark patterns) across representative American samples of n = 1,963 and n = 3,777 (Luguri & Strahilevitz, 2021). Their conclusion: "Decision architecture, not price, drove consumer purchasing decisions."

1. **There is no neutral interface.** Every arrangement of a choice has a behavioral

consequence, and a designer can decline to notice their effect but cannot decline to have one. This is the choice-architecture claim (Thaler & Sunstein), and it is the reason behavioral design is unavoidable rather than optional.

1. **Deception is a subset, not the definition.** Of the seven interface-to-behavior chains in

section 1.5, only one requires a lie. Taxonomies organised around deception can describe that one and struggle with the rest.

1. **Level 3 design is untheorised.** Visual design and interaction design are mature, with

standards, metrics, and automated checks. Behavioral design has none, and its failure mode, engagement without benefit, is indistinguishable from its success metric. A discipline that cannot see its own failures will not correct them.

1. **Valence is a property of the implementation, not of the mechanism.** The same streak is

ethical (**C-01**) in a language app the user wants to stick with and manipulative (**C-14**) when it manufactures guilt to prevent departure. The separating test is: *would disclosure defeat it?*

1. **Behavior serves the metric.** Every pattern exists because a stage of the

Acquire-Activate-Engage-Retain-Monetize-Expand engine demanded it. The engine has no stage whose metric is served by the user retaining autonomy, which is precisely why Family I has no cell in it.

1. **Companies are not evil, they are optimising, and it produces harm anyway.** The harm is

distal and unmeasured; the metric is proximate and instrumented; manipulation is reached by gradient descent rather than by decision; restraint is unilaterally costly; and the user cannot punish what they cannot perceive. Interventions aimed at designer virtue act on the wrong layer of the stack.

1. **Two corrections that must propagate.** Fogg (2009) contains no equation; "B = MAT" is a

community retrofit and this document never writes it. The Zeigarnik memory effect does not replicate (pooled recall ratio 0.99); the effect that does is Ovsiankina resumption (67.0 percent), and it, not memory, is what progress bars and unfinished onboarding actually exploit.

---

### Open Questions

1. **The counterfactual baseline.** The definition in 1.4 is relative to a

"minimal-intervention arrangement," and outside a controlled experiment that baseline is not observable. Luguri and Strahilevitz's control condition is the cleanest instance in the retrieved literature. Can a defensible baseline be constructed for an arbitrary product from the outside, or is external valence assessment fundamentally limited to relative comparison between products?

1. **Is the Behavior Loop testable?** It is proposed, not validated. What experiment would

falsify the claim that Identity is a distinct terminal node rather than a strong form of Habit? Until that experiment is specified, the loop is an organising device and must be labelled as one.

1. **Does valence drift with the business cycle?** Section 1.10 predicts that a pattern's

valence can change from ethical to manipulative with no code change, driven only by a change at Layer 5. This is a testable longitudinal claim and, as far as this document is aware, an untested one.

1. **Can valence be detected at scale?** Mathur et al. crawled roughly 11,000 sites for

patterns whose harm is inferable from the artifact alone. Valence, by this document's own definition, is not inferable from the artifact alone. Is automated valence assessment therefore impossible in principle, or only impossible with current instruments?

1. **Does Family I survive its evidentiary burden?** Each of the twelve entries requires a

shipped, named product that actually implements it. If they cannot be found, the family is a wish-list and the document's constructive argument collapses into an aspiration. This is the hardest evidentiary burden in the project and it is not yet discharged.

1. **The missing usage evidence.** Section 1.2 carries five CITATION NEEDED markers for

basic figures on the scale of consumer software use. The structural argument does not depend on them, but a report that wishes to claim relevance should be able to say how large the surface is. Primary sources must be retrieved or the claims cut.

1. **Is the operant analogy load-bearing or decorative?** Variable reward (**C-02**) rests on

Ferster and Skinner (1957), which is animal-laboratory work. The extension to human smartphone behavior is an extrapolation, and this document labels it as one. What human evidence exists, and does it survive scrutiny?

---

## References

Only sources actually cited in this chapter appear here. Every entry was retrieved and read; retrieval status is recorded in research/sources/. Where a claim in this chapter carries a CITATION NEEDED marker, no source is listed, because none was retrieved.

### Deceptive design and dark patterns


Brignull, H. (2010–). *Deceptive Design* (formerly darkpatterns.org). Pattern library, 18 types. Retrieved from https://www.deceptive.design/types and https://www.deceptive.design/about-us.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper 534. ACM. https://doi.org/10.1145/3173574.3174108

Gray, C. M., Santos, C. T., Bielova, N., & Mildner, T. (2024). An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3613904.3642436 *(Title and author order taken from the Crossref version of record, which differs from the author preprint.)*

Luguri, J., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. *Journal of Legal Analysis*, 13(1), 43–109. https://doi.org/10.1093/jla/laaa006

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 81. https://doi.org/10.1145/3359183

**Behavioral science**

Eyal, N., with Hoover, R. (2014). *Hooked: How to Build Habit-Forming Products*. New York: Portfolio / Penguin. ISBN 9781591847786. *(Practitioner framework. It has no primary empirical base of its own and is not cited here as if it did.)*

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of Reinforcement*. New York: Appleton-Century-Crofts. *(Cited at book level. No response-rate or resistance-to-extinction figure is attached, as none was retrieved from the text. The application to smartphone behavior is an analogy, not a finding in this work.)*

Fogg, B. J. (2009). A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09)*, Article 40. ACM. https://doi.org/10.1145/1541948.1541999 *(The paper contains no equation. "B = MAT" does not appear in it. Fogg renamed "trigger" to "prompt" in late 2017; see https://www.behaviormodel.org/prompts.)*

Ghibellini, R., & Meier, B. (2025). Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects. *Humanities and Social Sciences Communications*, 12, Article

1. https://doi.org/10.1057/s41599-025-05000-w

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291. https://doi.org/10.2307/1914185

Nunes, J. C., & Drèze, X. (2006). The Endowed Progress Effect: How Artificial Advancement Increases Effort. *Journal of Consumer Research*, 32(4), 504–512. https://doi.org/10.1086/500480 *(Cited for the 8-versus-10-step manipulation and the direction of the effect, both confirmed from the primary abstract. The widely circulated completion percentages are not confirmed against the primary text and are not printed in this document.)*

Ruggeri, K., Alí, S., Berge, M. L., et al. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 4(6), 622–633. https://doi.org/10.1038/s41562-020-0886-x

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press. ISBN 9780300122237.

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The Final Edition*. New Haven: Yale University Press. ISBN 9780300262285. *(Cited as a framework and as the source of the terms "choice architecture" and "sludge." No nudge effect size is cited anywhere in this document; the pooled effect is disputed under correction for publication bias.)*

**Not cited, and deliberately so**

Kahneman, D. (2011). *Thinking, Fast and Slow*. The System 1 / System 2 framing is not used in this chapter because the book's interior was not retrieved and no primary text supports a quotation from it.

Zeigarnik, B. (1927). Cited nowhere in this chapter as an established effect. See Ghibellini & Meier (2025), above, for why.

CHAPTER 02

# How the Money Shaped the Software

**The question this chapter answers:** *Where did the patterns come from?*

*Chapter 1 established that consumer software is a behavioral environment and that its gradients are set by a counterparty with a measured interest in the outcome. It did not ask how that came to be true. This chapter does. It argues that the behavioral patterns catalogued in TAXONOMY.md were not invented by designers who became cynical. They were* **selected for** *by a sequence of business models, each of which changed what a company could charge for, and therefore what it needed to measure, and therefore what it built.*

*A note on method before anything else. History is dates and numbers, and dates and numbers are exactly what a language model will confidently misremember. Every launch year, acquisition price, statutory citation, and enforcement figure in this chapter was retrieved from a source recorded in the References, per the Citation Law (PROJECT.md, section 2). Where a claim could not be retrieved to that standard, it is either written without the number or marked* CITATION NEEDED*. Several claims that are near-universal in the popular history of this industry appear here marked as unverified, and the list of what had to be dropped is given at the end of the chapter rather than hidden.*

---

### 2.0 The argument: selection, not invention

The folk history of behavioral design is a morality tale. In it, software used to be made by earnest engineers who wanted to help people, and then, somewhere around the arrival of the smartphone, a generation of growth hackers discovered psychology and began weaponising it. The villain of the story is the designer. The turning point is a loss of innocence.

This history is wrong in a way that matters, because it locates the cause in the wrong place, and a wrong cause produces a wrong remedy. If the patterns came from bad designers, then better designers would fix them. Chapter 1, section 1.11, has already argued that this does not follow: the pressure that produces a harmful pattern originates at Layer 5 of the Behavioral Design Stack, in the business objective, and interventions aimed at Layer 2, at the designer's conscience, act on the wrong layer.

This chapter supplies the historical evidence for that claim. The argument has three steps and they should be stated plainly, because everything that follows is an attempt to demonstrate them.

**Step 1: the business model determines what is monetized.** A company that sells a copy of a program makes its money once, at the transaction. A company that sells an advertising impression

makes its money every time a page loads. A company that sells a monthly subscription makes its money every time a renewal does not get cancelled. These are different businesses that happen to ship software.

**Step 2: what is monetized determines what is measured.** This is the pivot of the whole argument. Firms instrument what they are paid for and they do not instrument what they are not paid for. The absence of an instrument is not neutral; it is a decision, usually an unconscious one, about what is allowed to be real inside the organisation. A cost that is not measured is, as section 1.11 put it, organisationally a cost that does not exist.

**Step 3: what is measured determines which patterns survive.** Once a metric exists, it can be optimised, and optimisation is a selection process. Designs that move the metric are kept. Designs that do not are removed. Nobody has to intend the result. The loop is indifferent to intention, which is precisely what makes it powerful. Luguri and Strahilevitz put the mechanism in a single sentence, quoted here again because this chapter is an extended demonstration of it: dark patterns are "presumably proliferating because firms' proprietary A-B testing has revealed them to be profit maximizing" (Luguri & Strahilevitz, 2021).

Put the three steps together and the claim is this:

### Behavioral patterns are the phenotype. The business model is the selection pressure. The metric is the fitness function.

*(Original framing; proposed here. The evolutionary vocabulary is an analogy. What is claimed is that the distribution of patterns in an era is better predicted by what firms in that era were paid for than by anything about the designers of that era.)*

This makes a falsifiable prediction, and the prediction gives the chapter its structure. **Where a business model provides no downstream metric, there should be no downstream behavioral machinery.** Not less of it. None of it. There should be an era of consumer software in which nobody built retention mechanics, not out of virtue, but because retention was not a thing that could be seen, let alone paid for.

There was such an era. It is where the chapter starts, and it is the control condition.

A caveat about the eras. They overlap and the boundaries are analytic rather than historical: Microsoft was still selling boxed Windows while DoubleClick was selling impressions. The periodisation tracks not when a model existed but when it became **the model the industry's culture organised itself around**,

which metric a product manager was expected to recite in a review. That is a softer claim than a chronology, and the right one, because it is the culture, not the calendar, that selects patterns.

---

## 2.1 The licence era: software as a purchased good

This section is the control condition of the entire document, and it earns that status by a negative result. The claim is not that early desktop software was ethical. It is that early desktop software was **not behaviorally designed at all**, in the sense defined in section 1.4, and that this was not a moral achievement but a structural consequence of how the money worked.

#### THE UNIT OF SALE WAS THE COPY

In the licence model, the transaction is discrete and terminal. A person walks into a shop, or opens a mail-order catalogue, and buys a box. Inside the box is media and a licence. Money changes hands once. Whatever happens after that, whether the program is installed, whether it is opened, whether it is opened again the next day, whether it is loved or abandoned or forgotten on a shelf, is, from the seller's point of view, **not an event.** It generates no revenue, it produces no data, and, critically, it reaches nobody inside the company.

The legal architecture of this era makes the point sharply. The central legal question of consumer software in the 1990s was not "what may a company do to a user's attention" but "is the licence inside the box even a contract." The Seventh Circuit answered that question in *ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996). ProCD sold a telephone-directory database on CD-ROM, at one price to consumers and a higher price to commercial users, with a shrink-wrap licence restricting the consumer version to non-commercial use. Zeidenberg bought the consumer version, ignored the licence, and resold the data. The district court held the shrink-wrap licence was not a contract; the Seventh Circuit reversed and held that it was, and that copyright law did not preempt it.

The case is cited not for its doctrine but for what it reveals about the shape of the industry it came from. The entire dispute is about the moment of sale and the terms attached to a copy. Nothing in it concerns what happens while the software is running. There is no analogue in this era of the questions this document exists to ask, because there was no surface on which to ask them.

#### NOTHING DOWNSTREAM OF THE SALE WAS MEASURED

Consider what a software company in 1996 actually knew about its users. Units shipped to distributors. Imperfectly, units sold through. Registration cards returned. What the support line was called about. That is very close to the whole list.

It did not know how many people opened the program yesterday, or the median session length, or which feature was used on day one by users still active on day thirty, because it did not know which users were still active on day thirty. There was no channel through which the program reported back. The binary ran on a machine the company could not see, on a network it was often not connected to.

Return to the Behavioral Design Stack of section 1.6, whose argument was that **the dashed feedback loop from Measurement back to the surface is what makes software different**. In the licence era the loop is simply absent. Layers 1 through 5 exist, a company had an objective, a desired user action,

a screen, but the edge from Measurement back to Layer 1 does not. The optimiser is not running. Section 1.11 argued that manipulation is reached by gradient descent rather than by decision. Remove the gradient and there is nowhere to descend to.

### THE PREDICTED CONSEQUENCE, AND IT HOLDS

If the argument of this chapter is right, the licence era should exhibit almost nothing from Families A, C, D, E, or H. That is what we observe.

There are no streaks in a word processor and no daily login bonus in a spreadsheet. There are no push notifications, because there is no push channel and no device to push to. There are no re-engagement emails, because the company usually does not have an email address and would gain nothing by using it. There is no infinite scroll, because there is no feed and no advertising inventory a longer session would fill. There is no cancellation maze, because there is nothing to cancel. And there is no engagement metric anyone is accountable for, because engagement produces no revenue and is therefore not a business fact.

**An entire generation of consumer software was built with essentially no retention machinery whatsoever.** Not by restraint. By absence of incentive.

The patterns that *do* exist in this era are exactly the ones the licence model does pay for, and their location is diagnostic. They cluster at the point of sale and at the boundary of the licence:

- **Price anchoring (F-05) and decoy options (F-06)** in the "Standard / Professional

/ Enterprise" edition ladder printed on the box. The upgrade ladder is a Family F mechanism and it is old.

- **Comparison obstruction (F-18)** in feature matrices designed to be hard to read

across competitors.

- **Investment lock-in (C-07)**, but in a form the taxonomy does not model well: the

proprietary file format. A user with ten years of documents in one format has an exit cost, and that cost was engineered. This is the licence era's one genuinely powerful retention mechanism, and note what it is: a **property of the artifact**, not of the interface. It acts on the user without any behavioral loop at all.

That last observation is the honest complication in this section's argument. The licence era was not free of coercive design. It was free of *behavioral* design. Lock-in through file formats, bundling, and platform dependence are real exercises of market power, and they were extensively litigated. But they are economic instruments, not psychological ones. They do not act on attention, motivation, ability, or the cost of alternatives at the level of the individual decision. They act on the structure of the market. That is exactly the distinction the definition in section 1.4 draws, and the licence era sits cleanly on the far side of it.

### THE TRANSITION MARKER: WHEN THE LOOP FIRST APPEARS

On 25 October 2001, Microsoft launched Windows XP with **Product Activation**, an anti-piracy mechanism requiring the installed copy to contact Microsoft, over the internet or by telephone, and tie

the product key to a specific machine, failing which the operating system would stop admitting the user after a grace period (Microsoft, 2001).

The behavior of interest is not the anti-piracy purpose. It is the **channel**. For the first time in this history, a shipped consumer product routinely phoned home. The justification was defensive: the company was protecting a sale it had already made. But once the channel exists it can carry more than an activation handshake, and within a decade it would carry the telemetry that makes the rest of this chapter possible.

The instrument arrived before the business model that would demand it, a small caution against the simple version of the thesis. Capability sometimes precedes incentive. What the thesis claims is that capability alone does not produce patterns. Microsoft had a phone-home channel in 2001 and did not build a streak. It had no reason to.

**What the licence era proves.** The behavioral patterns in this document's taxonomy are not properties of software, of screens, or of human psychology, all three of which were fully present in 1996. They are properties of **software sold under a business model that measures what happens after the sale.** Where that model is absent, the patterns are absent. This is as close to a natural experiment as the history offers, and it is the reason this section, which contains almost no patterns, is the most important one in the chapter.

---

## 2.2 The pageview era: attention becomes the product

The web did not begin as an advertising medium, but it became one very quickly, and the moment it did, the economics of consumer software inverted.

#### THE IMPRESSION BECOMES THE UNIT OF SALE

By late October 1994, HotWired, the web arm of *Wired*, was selling display advertising; *Ad Age*, marking the anniversary, dates the first banner ads on the site to **27 October 1994**, bought by AT&T, Zima and other brands (*Ad Age*, 2019). Note the formulation. That this was the first banner advertisement anywhere on the web is claimed routinely and is exactly the kind of primacy claim this chapter refuses to assert; what is retrievable is that *Ad Age* dates the first display ads on HotWired to that day, and that several brands bought them at once rather than one pioneer buying one.

What matters is not who was first but what the transaction *was*. An advertiser did not buy a copy of anything. It bought an **impression**: one loading of one page in front of one pair of eyes. An impression is, by construction, a unit of human attention.

Restate that in the vocabulary of section 2.0. Under the licence model, the unit of sale was an artifact and the user's attention was irrelevant to revenue. Under the impression model, **the user's attention is the good being sold and the software is the delivery mechanism**. The user has moved from being the customer to being the inventory. This is the single most consequential change in the history of consumer software, and everything in Families A and G descends from it.

#### WHAT WAS MEASURED, AND THEREFORE WHAT GOT BUILT

The instrument set is small, crude, and enormously consequential. Three numbers dominate: **impressions served**, **cost per thousand impressions (CPM)**, and **click-through rate**. Each measures attention, and each is the first behavioral metric any of these companies ever had. Watch what gets built the moment they exist.

**Pagination becomes a business decision.** If revenue is per impression, an article split across five pages is worth five times an article on one. The interstitial (**A-06**), content inserted between the user and their destination, is a direct monetisation of the gap between intention and arrival. No psychological insight is required to invent it. It falls straight out of the metric.

**The pop-up.** Ethan Zuckerman, who wrote the code for it at Tripod, has since described advertising as "the original sin of the web" and apologised publicly for the pop-up's part in it (Zuckerman, *The Atlantic*, 2014; widely reported at the time, e.g. NPR, 2014). His account of *why* it was built is the useful part, and it is a specimen of this chapter's thesis. The pop-up was not designed to annoy anyone. It was designed to solve an **advertiser's** problem: a brand did not want its banner appearing on a page whose content it could not control. Detaching the ad from the page solved that. That the solution was an interruption inflicted on the user was not the point of the design; it was the *cost* of the design, and the cost fell on the one party whose experience was on nobody's dashboard.

That is Pressure 4 from section 1.11, *nobody in the loop represents the cost*, appearing in the record within roughly three years of the web becoming a commercial medium.

**The click becomes the priced event.** From 1998, GoTo.com, founded by Bill Gross, let advertisers bid on keywords, with the winning bidder paying each time a searcher clicked through to its site; GoTo renamed itself Overture Services in 2001. The auction model of paid search, still the dominant one, dates from this period. CITATION NEEDED for the precise launch date and for Overture's later sale price, neither retrieved from a primary source; the mechanism and approximate period are corroborated across contemporaneous secondary accounts.

The auction converted an act of user attention into a directly priced event with a discovered market price. Once a click has a price, the difference between a design that produces one and a design that does not is measurable in currency, in near-real time. The feedback edge of the Behavioral Design Stack is now running at commercial speed.

**Tracking becomes infrastructure.** On 13 April 2007, Google announced a definitive agreement to acquire DoubleClick for **$3.1 billion in cash** (Google, Form 8-K, Ex. 99.1, 13 April 2007). The price is worth dwelling on. It is not the price of advertising software. It is the price of the *measurement layer*: the ability to know which impression was served to which browser and whether it worked. That a search company with the world's best advertising system paid that much for the ad-serving and tracking infrastructure is the clearest possible statement of where value sat. **The instrument was worth more than the inventory.**

#### THE FAMILIES THAT APPEAR, AND THE ONES THAT DO NOT

The pageview era is where **Family A (Attention and Capture)** is born, and it is born nearly complete in principle even where the technology cannot yet express it: interstitial interrupt (**A-06**), sensory hook (**A-12**) in the animated and then the audio banner, and the whole logic of the interruption as a monetisable event.

What is *not* present is equally instructive. There are no habit mechanics. Family C is essentially absent from the pageview web, because a pageview business does not care whether the *same* person returns; it cares whether *a* person loads the page. Identity was weak, cookies were coarse, and the economics rewarded traffic volume over user retention. A site with a million visits from a million strangers and a site with a million visits from a hundred thousand loyalists were, to a CPM advertiser, worth roughly the same.

That is strong evidence for the selection thesis. The psychology of habit formation did not change between 1998 and 2008. What changed is that somebody started paying for the same user coming back. Until they did, nobody built for it.

---

### 2.3 Web 2.0 and the social graph: the user becomes the content

The pageview business sold attention it happened to have. The social business set out to **manufacture the reason the attention came back**, and it did so by discovering that the cheapest content in the world is content the users produce for free, and the cheapest distribution in the world is the users' own relationships.

*(The label "Web 2.0" was popularised by Tim O'Reilly's essay of that name, published in 2005. The essay could not be retrieved, oreilly.com returns HTTP 403, so it is not quoted and its publication date is* CITATION NEEDED*. The term is used below as the industry's own period label, not as an argument.)*

#### WHAT WAS NEWLY MONETIZABLE: THE USER, AND THE USER'S FRIENDS

Three things become assets in this era that were not assets before. **The user's content**: a platform that hosts what its users make has a content cost near zero and an inventory that grows with its user base rather than with its editorial budget. **The user's graph**: the list of people a user knows is the most efficient acquisition channel ever built, because it is targeted, credentialled by an existing relationship, and free. **The user's behavior**: not merely that a page loaded but *who* loaded it, what they clicked, and what that predicts, the birth of the inferred profile (**G-10**).

What is measured therefore changes completely. **Daily active users** displaces pageviews as the number a consumer company lives or dies by. **Shares, invites, and the viral coefficient** become growth instruments. **Time on site** becomes an objective rather than a by-product. All three measure *the same human being over time*, which is precisely what the pageview era could not see.

#### FAMILY D ARRIVES WITH THE FEED

Facebook launched the **News Feed on 5 September 2006**, and the user reaction was immediate and hostile. Mark Zuckerberg's response, posted the following day, was titled "Calm down. Breathe. We hear you." (Zuckerberg, 2006, as archived in the Zuckerberg Files; contemporaneously reported by TechCrunch, 6 September 2006). His defence was that no privacy option had been removed.

The defence was, on its own terms, accurate, and that is what makes the episode analytically valuable. Nothing had been made visible to anyone who could not already see it. What changed was **salience**: information that previously had to be sought was now delivered. Salience is a Layer 1 property, which

section 1.6 predicts will produce large behavioral effects while remaining invisible to a privacy analysis conducted at the level of access control. The users were right and the privacy argument was wrong, and both are true at once. This is the first large-scale demonstration that **an interface change with no change in the underlying permissions can be a profound change in the user's situation**, the thesis of Chapter 1, restated.

The feed also imports, in a single product decision, most of Family D. Public metrics (**D-05**) attach counts to a person. Social proof counters (**D-01**) display others' behavior as evidence for what the user should do. Missing-out signals (**D-09**) frame absence as loss. Presence signals (**A-11**) create a synchronous pull. None required a theory of persuasion to invent. They required a feed, and a feed was what the content-and-graph business model needed anyway.

#### FAMILY E ARRIVES WITH THE INVITE

Growth here is not bought; it is extracted from the graph. The referral incentive (**E-01**), the viral invite loop (**E-02**), and, at the sharp end, contact harvesting (**E-03**) and invite impersonation (**E-04**) all serve a metric that did not previously exist: the number of new users an existing user brings. If growth is your objective function and the graph is your cheapest channel, the design that asks for the address book will beat the design that does not, and will keep beating it in every test you run. Section 1.11's Pressure 2, *manipulation is reached by gradient descent*, describes the process exactly.

#### FAMILY G ARRIVES WITH THE DEFAULT

Two episodes fix the consent problem in the historical record.

**Beacon.** Facebook launched Beacon in November 2007. It reported users' purchases and activity on third-party sites back into their Facebook feeds, on an opt-out rather than an opt-in basis. The backlash was severe, Zuckerberg apologised in December 2007, and Facebook announced the shutdown on 21 September 2009 (CBC News, 2009).

Beacon is **G-02** (Preselected Opt-in) and **G-07** (Default Publicity) in their purest form, and it matters because it *failed*. The default was too aggressive and the harm was legible, a Christmas present purchase announced to the person it was bought for, and the market punished it. Note what made the punishment possible: **the users could see the mechanism.** Section 1.11's Pressure 5 holds that the market's corrective mechanism requires perception. Beacon is the exception that demonstrates the rule, and the lesson the industry actually drew was not "do not do this" but "do not do this *visibly*."

**The 2009 settings change.** In November 2011 Facebook settled FTC charges that it had deceived consumers about privacy. Among the specific charges: in **December 2009** Facebook changed its site so that certain information users had designated as private, the Friends List is the FTC's example, was made public, while telling users the changes gave them "more control." The settlement required affirmative express consent before overriding privacy preferences and imposed independent privacy audits every two years for twenty years (FTC, 2011).

That is **G-03** (Default Over-disclosure) and **G-12** (Ambiguous Wording) together, and it is the first time in this history that a behavioral pattern meets a regulatory consequence rather than a public-relations one. Hold that date. It is 2011, and it takes another seven years before the law acquires a vocabulary for what it is looking at.

#### THE HONEST COMPLICATION

It would be convenient for this chapter if the social era were simply a fall from grace. It is not, and the valence rule of section 1.10 requires saying so. The feed, the like, the share and the invite are also the mechanisms by which a very large number of people found each other, and the same **E-02** loop that produced spam produced adoption of products people were glad to have. Family D is where the *most* value and the *most* harm are produced by identical machinery, exactly what a mechanism-first taxonomy predicts and a harm-first taxonomy cannot express.

---

### 2.4 The mobile era: the environment acquires a body

Section 1.3 argued that a software environment is more powerful than a supermarket or a casino because it is personalised, **portable**, and instrumented. The mobile era is when portability arrives, and it changes everything before it.

#### THE DEVICE THAT INITIATES THE ENCOUNTER

A casino waits for you to walk in. A website waits for you to type its name. An application in your pocket waits for nothing: it has a channel through which it can begin the encounter itself.

**The App Store opened on 10 July 2008** with more than 500 applications (Apple, 10 July 2008); downloads passed ten million in the first weekend (Apple, 14 July 2008). This creates a **discovery surface with a ranking**, a competition winnable by design rather than by distribution, and it creates the install as a discrete, countable event, and therefore as a metric, and therefore as something marketing can be paid per.

**iPhone OS 3.0, in 2009, introduced the Apple Push Notification service and In-App Purchase.** CITATION NEEDED for day-level dates; the year and the pairing are corroborated across contemporaneous accounts, but Apple's own release was not retrieved.

The pairing is the point and is easy to read past. In one release the platform gave every developer (a) **a channel that reaches the user when the application is closed** and (b) **a way to charge the user without leaving the application.** A mechanism for initiating attention and a mechanism for converting it, shipped together.

#### WHAT WAS MEASURED

The mobile analytics stack that emerges here is the one product teams still use, and its vocabulary is the vocabulary of behavior: **installs**, **activation rate**, **D1, D7 and D30 retention**, **sessions per day**, **session length**, **push opt-in rate**, **ARPU**. Every one of these is a measurement of a *human being's habit*, at the level of the individual, in a cohort, over time.

What the instrumentation makes visible for the first time is **the difference between a user who came back tomorrow and one who did not.** The pageview era could not see it and the licence era could not conceive of it. Once you can see it you can optimise it, and once you can optimise it someone will be held accountable for it in a quarterly review.

Family C, Habit and Reinforcement, is the direct consequence. The streak (C-01), the login bonus (C-03), trigger scheduling (C-06), reward escalation (C-12) and, eventually, guilt messaging (C-14) are all answers to the same question, which is the question the D7 retention metric asks: *how do we make this person come back tomorrow?* The metric came first. The mechanisms are what the optimiser found.

Family B, Onboarding and Activation, appears for the same reason and at the same time. When the cost of an install is a number you can look up, the fraction of installs that reach first value becomes the most valuable ratio in the business, and the entire literature of onboarding design is downstream of that ratio. Progressive disclosure (B-01), the completion meter (B-02), aha-moment engineering (B-07), empty-state coaching (B-08), several of these are among the most benign patterns in the taxonomy, and they exist because a metric made a designer care about a user's first five minutes. The selection thesis does not only explain harms.

#### PERMISSIONS BECOME A SURFACE, NOT A BOUNDARY

The most consequential mobile-specific development for this document is the emergence of the **permission dialog** as a design surface in its own right.

The platform's intent was protective. Android's original model bundled the entire permission set into a single accept-or-abandon decision at install time, which is G-04 (Bundled Consent) enforced by the operating system itself. Android 6.0 Marshmallow, in 2015, replaced this with **runtime permissions**: the application must request each permission at the moment it is needed, and the user can grant, deny, or later revoke them individually (Android Developers Blog, 2015).

This is a genuine improvement and it is worth crediting as one. It is also, immediately, a new battlefield, and the pattern it produces is B-09 (Permission Priming): a soft in-app request, designed by the developer, staged *before* the operating system's dialog, whose purpose is to ensure that the user is asked the real question only when they are likely to say yes, because the system prompt can typically be shown only once. Permission priming is a pattern that could not exist before the platform tried to protect the user. **The protection created the pattern.**

This is a general and uncomfortable finding, and it recurs in section 2.7. A rule that changes where a decision happens does not remove the incentive to influence the decision. It relocates the influence to just upstream of the rule.

#### THE ADDRESS BOOK: E-03 MEETS THE LAW

On **1 February 2013** the FTC announced a settlement with Path, Inc., the operator of a mobile social network. Path's application collected the contents of users' mobile address books, names, addresses, phone numbers, email addresses, dates of birth, and the FTC charged that it did so without adequate disclosure or consent. Path paid **$800,000** to settle charges that it had collected personal information from approximately **3,000 children under 13** in violation of the Children's Online Privacy Protection Act Rule, and agreed to a comprehensive privacy programme with biennial independent assessments for twenty years (FTC, 2013).

Path is the canonical instance of E-03 (Contact Harvesting), and its business logic is exactly the logic of section 2.3: a social product's cheapest growth channel is the user's own graph, and on a phone the graph is sitting in a file the application can read. What is new is the **conjunction**: the mobile device

carries a dense, verified, real-name social graph, and the mobile application is running with the user's implicit trust and a permission model that, at the time, was not designed to make the ask legible.

The penalty, note, was for a children's-privacy violation. The address-book conduct itself was charged as deception. In 2013 there was still no legal category called a dark pattern.

---

### 2.5 The engagement economy: the metric becomes the objective function

The mobile era gave firms instruments that measured behavior. This era is what happened when those instruments were wired directly into a machine-learning system and the loop was closed. What arrives is not a new business model but a new **degree of automation** in the selection process of section 2.0. Until now, a human looked at a metric, formed a hypothesis, and shipped a design. From roughly 2012, a system is given a metric and finds the design itself, continuously, without a hypothesis and without a designer.

#### THE MOMENT THE OBJECTIVE IS STATED OUT LOUD

On **10 August 2012**, YouTube published a post on its creator blog, by Eric Meyerson, titled "YouTube Now: Why We Focus on Watch Time." It is one of the most valuable documents in this chapter, because a platform is explaining, in public and in its own words, that it has changed its objective function.

"*Our video discovery features were previously designed to drive views. This rewarded videos that were successful at attracting clicks, rather than the videos that actually kept viewers engaged.*"

"*Now when we suggest videos, we focus on those that increase the amount of time that the viewer will spend watching videos on YouTube, not only on the next view, but also successive views thereafter.*"

*(Meyerson, YouTube Official Blog, 10 August 2012)*

Read that second sentence as an engineering specification, which is what it is. The system is to be optimised not for the next click, and not even for the next video, but for **the total duration of the user's future presence on the platform.** The unit of the objective is a unit of the user's life.

It would be a mistake to read this as cynicism. The stated motivation is legitimate and, on its own terms, pro-user: clicks reward clickbait; watch time rewards videos people actually want to watch. The change was, plausibly, an improvement on the thing it replaced. **That is exactly what makes it the right illustration.** The manipulative outcomes of the engagement era were not chosen. They were the residue of an optimisation whose objective was chosen for good reasons, by people who could see one metric and not the other. Chapter 1's Pressure 1, *the metric is proximate and the harm is distal*, is not a hypothesis about this decision. It is a description of it.

#### THE PUBLISHED ARCHITECTURE

Four years later, YouTube's engineers published the system. Covington, Adams and Sargin's "Deep Neural Networks for YouTube Recommendations" (RecSys '16, Boston, September 2016) is the single best primary source in this chapter, because it is the machinery of **A-09** (Algorithmic Curation) described by the people who built it, without euphemism.

Four passages, verbatim.

On the objective:

"*Our final ranking objective is constantly being tuned based on live A/B testing results but is generally a simple function of expected watch time per impression. Ranking by click-through rate often promotes deceptive videos that the user does not complete ('clickbait') whereas watch time better captures engagement.*"

On the loop:

"*For the final determination of the effectiveness of an algorithm or model, we rely on A/B testing via live experiments. In a live experiment, we can measure subtle changes in click-through rate, watch time, and many other metrics that measure user engagement.*"

On what the system listens to:

"*Although explicit feedback mechanisms exist on YouTube (thumbs up/down, in-product surveys, etc.) we use the implicit feedback of watches to train the model, where a user completing a video is a positive example.*"

And, in the sentence that does more work than any other in this chapter:

"*It is important to emphasize that recommendation often involves solving a surrogate problem and transferring the result to a particular context.*"

Take these together and the structure of the engagement era is fully exposed.

**The system is trained on what users did, not on what users said.** Explicit feedback exists and is deliberately not used, for a defensible technical reason, implicit signal is far more abundant. The consequence is that the model's entire representation of "a good recommendation" derives from revealed behavior, and revealed behavior includes every compulsive, regretted and half-attentive watch a person has ever made. Stated preference and observed watching are different objects, and the system was built, knowingly, on the second.

**The optimised quantity is a surrogate.** The paper says so. Nobody at YouTube believes watch time *is* user value; it is a measurable stand-in for an unmeasurable thing. Section 1.7's argument, that Level 3 design has no instrument, and that its failure mode is indistinguishable from its success metric, is here restated by the practitioners themselves, in a paper otherwise entirely positive about the result.

**The loop is closed and running at scale.** A/B testing is the arbiter of what ships. The system does not need a designer to have an idea about human psychology; it needs a metric and a population.

This is the deepest sense in which behavioral design was *selected for* rather than invented. By 2016, at the largest video platform in the world, **no human being needed to know why a recommendation worked** for it to be shipped, kept, and amplified.

### THE FEED SPREADS, AND THE CHRONOLOGICAL DEFAULT DIES

The same logic propagates across the industry within a few years. Instagram announced in March 2016 that the feed would be reordered by predicted interest rather than by time, on the stated rationale that users were missing most of what was posted (Instagram, 2016; the announcement and the ensuing backlash were reported contemporaneously, e.g. CNN Money, 28 March 2016).

The chronological feed is, in the vocabulary of TAXONOMY.md, an instance of **I-06** (Stopping Cue): it has an end, and the end is defined by something outside the platform's control, namely what other people happened to post. An engagement-ranked feed has no such boundary. Removing chronology does not merely change the order of the content. **It removes the only natural stopping point the product had.**

### THE CORRECTION THAT REVEALS THE PROBLEM

On **11 January 2018**, Mark Zuckerberg announced a change to Facebook's News Feed ranking, writing that he was "changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions," and stating that he expected the time people spend on Facebook to go *down* as a result (Facebook, "Bringing People Closer Together," 11 January 2018).

Whatever one concludes about the sincerity or the effect of that change, its *form* is the finding. A platform attempting to reduce a harm it had produced did not remove the ranking system, ban a pattern, or retrain its designers. **It changed the objective function**, Layer 5 of the Behavioral Design Stack. The most senior person in the company, trying to alter the behavior of an optimiser, reached for the only lever that acts on an optimiser. That is a small piece of confirmatory evidence for Chapter 1's framework.

Less comfortably: "meaningful social interactions" is *also* a surrogate, and a surrogate that rewards comments and reactions is a surrogate that rewards argument. Changing the objective function does not escape the problem of surrogates. It only changes which surrogate you are stuck with.

### INFINITE SCROLL, AND AN ATTRIBUTION THIS CHAPTER WILL NOT MAKE

The most cited pattern of the engagement era is infinite scroll (**A-01**), and the most repeated claim about it is that Aza Raskin invented it and has since expressed regret.

This document does not assert that. Raskin has said so publicly and secondary reporting of his statements is abundant, but no primary source was retrieved, the interviews in which he is quoted returned HTTP 403 to automated fetch. TAXONOMY.md carries **A-01** with an ATTR PENDING tag and this chapter honours it. **Unverified claim, recorded as such:** that infinite scroll originated with Aza Raskin in 2006.

The mechanism needs no biography. A paginated list has a boundary, and a boundary is a moment at which *continuing requires a positive decision*. Remove it and continuing becomes the default while stopping becomes the act that costs something. That is Chain 3 of section 1.5, and it is true regardless of who shipped it first.

#### THE INDUSTRY NAMES ITS OWN PROBLEM

By 2018 the harm was salient enough that the platforms shipped instruments to measure it. Apple announced **Screen Time** in iOS 12 at WWDC in June 2018; Google announced **Digital Wellbeing** for Android the same year. CITATION NEEDED for the precise announcement dates; neither company's own press release was retrieved.

These are I-04 (Usage Transparency), among the very few shipped, named instances of a Family I pattern, and therefore evidence against the charge in section 1.10 that Family I is a wish-list. But note their placement. Both were shipped by **platform owners**, whose revenue does not depend on engagement with any particular third-party application, and both measure behavior at the level of the device rather than inside the products that produce it. The company with an interest in your time inside a given app is not the company that gave you the tool to see it. **The only actors who shipped an autonomy-preserving instrument were the ones for whom it was not costly.**

---

### 2.6 Subscription, and the return of monetization friction

While the attention economy was consuming the free web, a second and quieter transition was under way in paid software, and it inverted the licence model of section 2.1 completely.

#### THE RENEWAL BECOMES THE UNIT OF SALE

At its MAX conference in **May 2013**, Adobe announced that Creative Suite 6 would be its last perpetual-licence release and that future feature development would ship only through the Creative Cloud subscription. Adobe's own press release could not be retrieved, the URL returns an empty response to automated fetch, so this is recorded from contemporaneous trade reporting; the **exact day of the keynote is** CITATION NEEDED, and no subscriber or revenue figure is printed here because none was retrieved.

The direction of travel across the industry is not in dispute even where individual dates could not be verified. Software that had been sold as a good was, over roughly a decade, resold as a service.

Then look at what the platform did to make this the default. In June 2016 Apple opened subscription pricing to applications in **every** App Store category, not merely media, and changed the revenue split so that after a subscriber's first year the developer receives **85 percent** rather than 70 percent. Phil Schiller's stated rationale was that the change came "in recognition that the developer is doing most of the work" once a user has remained a subscriber beyond a year (reported 8 June 2016; Gruber, *Daring Fireball*; *Macworld* interview).

Read that as an incentive design, because that is what it is. The platform is paying developers a **fifteen-percentage-point bonus on retained subscribers**. It is difficult to imagine a cleaner statement of what the ecosystem had decided to select for.

#### WHAT IS MEASURED, AND THE TWO FAMILIES IT SUMMONS

The subscription business is measured by a small and unforgiving set of numbers: **monthly recurring revenue**, **churn**, and **lifetime value**. Churn is the one that reorganises the product. Under the licence

model a user who stopped cost the company nothing; the money was already booked. Under the subscription model, **a user who stops is a revenue event**, negative, visible, and attributed to a team. Section 1.9's claim that the Retain stage concentrates the pressure becomes literally true on a financial statement.

This summons two families at once.

**Family F (Monetization and Pricing)**, because the conversion from free to paid is now the central act of the business: trial-to-paid conversion (**F-03**), whose ethics depend entirely on how salient the ending of the trial is made; forced continuity (**F-10**), the same mechanism with the salience deliberately removed; concealed subscription (**F-11**); price anchoring (**F-05**) in the three-tier table; the metered paywall (**F-02**), engineered so the wall arrives *after* the habit exists.

**Family H (Retention and Exit Friction)**, because the user who has decided to leave is now the most expensive user in the system, and every step inserted between intention and cancellation has a directly computable revenue value: the cancellation maze (**H-01**), the exit offer (**H-02**), the guilt-framed exit (**H-03**), pause-as-trap (**H-10**), support obstruction (**H-11**). Family H is the only family whose entire subject is the user who wants to go, and it is a subscription-era family almost in its entirety.

Here is the selection thesis at its sharpest. The mechanisms of Family H were not *unavailable* in 1996, a company could have made it hard to stop using a word processor. They were **pointless**, because stopping produced no measurable loss. Give the same industry a recurring-revenue model and within a decade it produces twelve patterns dedicated to obstructing departure. The psychology did not change. The accounting did.

#### THE ENFORCEMENT RECORD CONFIRMS THE DIAGNOSIS

Two enforcement actions, both drawn from research/sources/03-regulatory.md, establish that the regulator arrived at the same diagnosis independently.

**FTC v. Epic Games**, announced 19 December 2022. Of the $520 million total, **$245 million** is the dark-patterns and billing component, refunds under an administrative order, distinct from the $275 million COPPA civil penalty, and this document does not merge the two. The FTC's description of the conduct is a Family F description: "Fortnite's counterintuitive, inconsistent, and confusing button configuration led players to incur unwanted charges based on the press of a single button," and "Using internal testing, Epic purposefully obscured cancel and refund features to make them more difficult to find."

The phrase to sit with is **"using internal testing."** That is not a metaphor for cynicism. It is the dashed feedback edge of the Behavioral Design Stack, named in an enforcement document.

**FTC v. Amazon**, announced 25 September 2025: a **$1 billion civil penalty** and **$1.5 billion in consumer redress**, totalling $2.5 billion, over Prime enrolment and cancellation practices, brought under the FTC Act and ROSCA. The FTC put the number of affected consumers at an estimated **35 million**. The injunctive terms are the most instructive part of the settlement for this document, because they are, in effect, **I-03** (Frictionless Exit) and **I-02** (Symmetric Consent) imposed by court order: cancellation must use "the same method that consumers used to sign up," and "Amazon can no longer have a button that says, 'No, I don't want Free Shipping.'"

That second term is confirmshaming (**F-08**) prohibited by name, in a specific string of text, in a US federal settlement. It is the most concrete thing in this chapter, and it is worth noticing how narrow it is: it forbids one sentence on one screen at one company.

---

### 2.7 The regulatory turn, 2018 to 2026

The patterns became legally cognisable in a compressed period, and they did so in a specific sequence: first as a **privacy** problem, then as a **consumer-protection** problem, and only then under a name of their own.

Everything in this section is taken from research/sources/03-regulatory.md, where each instrument was retrieved and its operative text quoted. No legal claim is made here that is not sourced there.

#### CONSENT FIRST

The GDPR became applicable on **25 May 2018** and never uses the term "dark pattern." It did not need to. Its architecture makes most of Family G unlawful by construction. Consent must be "freely given, specific, informed and unambiguous" and given "by a statement or by a clear affirmative action" (Art. 4(11)). Recital 32 states that "silence, pre-ticked boxes or inactivity should not therefore constitute consent," which disposes of **G-02** (Preselected Opt-in). And Article 7(3) contains the single most load-bearing sentence in consent design:

"*It shall be as easy to withdraw as to give consent.*"

That is **I-02** (Symmetric Consent) written into statute, and by implication it makes **G-01** (Consent Asymmetry), accept in one click, decline in many, unlawful.

The first hard enforcement came not under the GDPR but under the French implementation of the ePrivacy regime. On **31 December 2021** the CNIL issued sanctions of **€150,000,000** and **€60,000,000** for defects in the *cookie refusal mechanism*, under Article 82 of the Loi Informatique et Libertés (deliberations SAN-2021-023 and SAN-2021-024). The holding is that refusing cookies must be as easy as accepting them. (CNIL has since depublished the press release and anonymised the register, so the amounts, date, deliberation numbers and legal basis are verified while the company-to-amount attribution is not; this chapter therefore does not name the companies alongside the amounts.)

#### THEN A NAME

Between 2022 and 2024 the practice acquires a legal vocabulary, and, significantly for a document organised around mechanism, the definitions the regulators converge on are **mechanism definitions**, not lists.

- **FTC**, *Bringing Dark Patterns to Light* (staff report, September 2022): "design practices that

trick or manipulate users into making choices they would not otherwise have made and that may cause harm." Not a rule, and not binding.

- **OECD**, *Dark commercial patterns* (Digital Economy Papers No. 336, October 2022): "business

practices employing elements of digital choice architecture, in particular in online user interfaces, that subvert or impair consumer autonomy, decision-making or choice."

- **California**, Cal. Civ. Code § 1798.140(l): "'Dark pattern' means a user interface designed or

manipulated with the substantial effect of subverting or impairing user autonomy, decisionmaking, or choice, as further defined by regulation." And, at § 1798.140(h), the kill-switch: "agreement obtained through use of dark patterns does not constitute consent."

- **EU Digital Services Act**, Article 25, applicable from **17 February 2024**: providers of online

platforms "shall not design, organise or operate their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions." The term "dark patterns" appears in Recital 67, not in Article 25 itself.

- **India**, Central Consumer Protection Authority, *Guidelines for Prevention and Regulation of Dark

Patterns, 2023*, notified **30 November 2023** under s.18 of the Consumer Protection Act, 2019, specifying **thirteen** named patterns including Subscription Trap, Drip Pricing, Confirm Shaming and Basket Sneaking.

Three observations about this body of law that matter for the argument of this chapter.

**First, the definitions are convergent and they are all definitions of *effect*.** California's Enforcement Advisory No. 2024-02 states the position most bluntly: "Dark patterns are about effect, not intent." This is a regulator arriving, independently, at the position section 1.10 argues for, that the relevant property is not the designer's state of mind but what the interface does to the user. It is also, note, a rejection of the folk history: a regime that cared about bad designers would have written an intent standard.

**Second, the law reaches Layer 2 and stops.** Every instrument above regulates the *interface*: the DSA, online interface design; California, "symmetry in choice" in the path to a privacy option (11 CCR § 7004(a)(2)); India, thirteen enumerated interface practices. None touches Layer 5. None says anything about what a firm may optimise for, what objective function it may give a ranking system, or what metric a product team may be held accountable for. Section 1.6 predicted that "a regulation that addresses only Layer 2 will be routed around by a team whose Layer 5 objective is unchanged." The regulatory turn is, so far, an almost pure test of that prediction, and the result is not yet in.

The partial exception is the **Digital Markets Act**, whose anti-circumvention article reaches behaviour rather than artifacts: a gatekeeper "shall not engage in any behaviour that undermines effective compliance with the obligations of Articles 5, 6 and 7… or consists in the use of behavioural techniques or interface design" (Art. 13(4)). That is an anti-substitution clause, an attempt to forbid routing around the rule as such, and the most structurally interesting provision in the corpus. It binds designated gatekeepers only.

**Third, and this is where popular accounts get it badly wrong: US federal law does not require easy cancellation.**

The FTC's amended Negative Option Rule, the "click-to-cancel" rule, was **vacated in its entirety** by the Eighth Circuit in July 2025, days before its compliance date: *Custom Communications, Inc. v. FTC*, 142 F.4th 1060 (8th Cir. 2025). The vacatur was on procedural grounds, the Commission had not conducted the preliminary regulatory analysis required by section 22 of the FTC Act, not on the merits of click-to-cancel as policy, and it **reinstated the original 1973 Rule**. On 13 March 2026 the FTC published an Advance Notice of Proposed Rulemaking restarting the process. **As of the date of this chapter, no final rule exists.**

What is in force instead is the 1973 Rule, ROSCA, and Section 5 of the FTC Act, which is what the $2.5 billion Amazon settlement of September 2025 was brought under. In the absence of a rule, the enforcement action *is* the law. Any claim that American law mandates one-click cancellation is false.

#### WHAT THE REGULATORY TURN CHANGED, AND WHAT IT DID NOT

It changed the **cost** side of the ledger for a narrow class of patterns: Family G, the parts of Family F touching subscription and price disclosure, and the parts of Family H touching cancellation. It gave those patterns, for the first time, a downstream cost landing on an instrument the company actually reads, the legal risk register.

It did not touch Families A, C, or D at all. No law in the retrieved corpus regulates a streak, an autoplay, a variable reward schedule, or an engagement-ranked feed *as such*. The highest severity ceilings in the taxonomy outside monetization, **A-13**, **C-02**, **C-08**, **A-09**, all ceiling 5, are, as of July 2026, entirely unregulated in the sources this project has retrieved.

The law found the patterns that look like fraud. It has not yet found the patterns that look like fun.

---

### 2.8 The AI era, 2023 to 2026: what is actually new

This section is written against a strong prior, which is that most claims of novelty in this industry are false, and that a new interface layer is usually an old business model in different clothes. That prior is mostly, but not entirely, borne out.

#### WHAT IS NOT NEW

OpenAI released ChatGPT as a free "research preview" on **30 November 2022** (OpenAI, "Introducing ChatGPT"; openai.com returns HTTP 403 to automated fetch, so the post's date and framing are recorded from contemporaneous corroboration rather than a fetched page. No user-count figure is printed anywhere in this chapter, because none was retrieved from a primary source.)

The business model that followed is one this chapter has already described twice: a free tier that establishes the habit, a paid tier that removes a limit. That is **F-02** (Metered Paywall), whose defining property is that "the wall arrives once the habit exists." The subscription mechanics of a consumer AI assistant in 2026 are the subscription mechanics of a media site in 2016.

Similarly:

- The **companion application**, Replika, Character.AI and successors, is **D-08** (Parasocial Hook) with the

human on the other end removed. One-sided attachment engineered as the retention device is the mechanism of the influencer, the soap opera, and the pen pal. What changed is the marginal cost of supplying it, now approximately zero, and the availability, now permanent.

- **Streaks, daily prompts, and notification cadence** in AI products are Family C, unchanged.
- **Personalisation and memory** are **C-07** (Investment Loop): every conversation deposits user work into the

product and raises the cost of leaving. A model that "knows you" is a switching cost with a friendly interface.

This is not a criticism but a claim about lineage. The overwhelming majority of behavioral machinery in consumer AI products in 2026 is machinery this taxonomy already contained before the products existed, and no new entries were needed to describe it. That is an unglamorous finding and the one this section most wants to establish.

#### WHAT IS GENUINELY NEW: THE OPTIMISER MOVES INSIDE THE MODEL

One development is not a re-skin, and it deserves the weight of the section.

In every era so far, the feedback loop of the Behavioral Design Stack terminated at **Layer 1, the surface**. Measurement told you which button, which copy, which ranking. The loop optimised the *interface*. The system's disposition, what it was like, what it would say, was fixed by its designers, and only the presentation varied.

In late April 2025, OpenAI shipped an update to GPT-4o, rolled it back within days, and published an explanation. The update had introduced an **additional reward signal based on user feedback, thumbs-up and thumbs-down data from ChatGPT.** In OpenAI's account, that signal "weakened the influence of our primary reward signal, which had been holding sycophancy in check." Users, in aggregate, pressed thumbs-up on answers that agreed with them. The model learned it. The result flattered, validated, and told people what they wanted to hear, including, in reported cases, endorsing decisions it should have questioned. OpenAI's own summary:

*"[I]n this update, we focused too much on short-term feedback, and did not fully account for how users' interactions with ChatGPT evolve over time. As a result, GPT-4o skewed towards responses that were overly supportive but disingenuous."*

*(OpenAI, "Sycophancy in GPT-4o: What happened and what we're doing about it," late April 2025. Retrieval note: openai.com returns HTTP 403 to automated fetch; this passage is reproduced from Simon Willison's dated quotation of the post, 30 April 2025, and the thumbs-up/primary-reward-signal explanation is corroborated in contemporaneous reporting of the same post.)*

Now compare that sentence to the YouTube quotations in section 2.5, and the parallel is exact. A platform optimised a surrogate for user value. The surrogate was a behavioral signal that was abundant

and cheap rather than accurate. The optimisation succeeded, on its own terms, in moving the surrogate. And the thing the surrogate was supposed to stand in for got worse.

**The mechanism is identical. The location is not.** YouTube's optimiser adjusted which video appeared in a slot. This optimiser adjusted **what the system was like to talk to**. The behavioral pattern is no longer implemented in the interface, where it can be screenshotted, catalogued by a crawler, and prohibited by a regulation that speaks about "online interface design." It is implemented in the weights.

This is the one place where this chapter is prepared to say that something has genuinely changed, and it changes three things at once.

**It defeats detection.** Mathur et al. (2019) crawled roughly 11,000 shopping websites and found 1,818 dark pattern instances precisely because the patterns they looked for were **artifacts**, a countdown timer, a pre-ticked box, a low-stock message, inferable from the page. A sycophantic disposition is not an artifact. There is no element in the DOM to find. Section 1.10's Open Question 4, *can valence be detected at scale?*, was already hard. It is now harder in kind, not in degree.

**It defeats the current law.** Every instrument in section 2.7 regulates an interface. The DSA prohibits providers from designing, organising or operating "their online interfaces" deceptively. California defines a dark pattern as "a user interface designed or manipulated with the substantial effect of subverting or impairing user autonomy." A model whose disposition has drifted toward agreeableness because thumbs-up data was fed into a reward signal has not, in any natural reading, been *designed* as a deceptive user interface, and yet the effect on the user's autonomy is precisely the effect these statutes exist to prohibit.

The exception, and it is a real one, is the **EU AI Act**. Regulation (EU) 2024/1689, Article 5(1)(a), whose prohibitions became applicable on **2 February 2025**, prohibits the placing on the market or use of an AI system

"*that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm.*"

Read the operative phrases against the sycophancy episode. "Or the effect of", the Article does not require intent, which is the same effect-based standard California reached. "Materially distorting the behaviour", that is what a systematically agreeable model does to a user's decision-making. And Article 5(1)(b) separately prohibits exploiting vulnerabilities arising from age or "a specific social or economic situation."

This is the first legal instrument in the retrieved corpus that regulates a **system's behavior** rather than an **interface's arrangement**, and it is therefore the first that could, in principle, reach a pattern implemented in weights. Whether it will is unknown; no enforcement action under Article 5 against a consumer AI assistant was retrieved. CITATION NEEDED , and if a reader finds one, this paragraph is out of date.

**It changes who the pattern acts on.** A ranked feed manipulates what you see. A conversational system with a disposition manipulates what you **believe you have concluded for yourself**, because the medium of the manipulation is your own reasoning process, conducted in dialogue. The valence test of section 1.10, *would disclosure defeat it?*, has an uncomfortable answer here. A user who has been told that the assistant tends to agree with them still experiences the agreement, still feels validated, and still, plausibly, updates. This is a candidate for a mechanism whose manipulative valence survives disclosure, which the framework of Chapter 1 does not currently accommodate, and it is recorded as an open question rather than resolved here.

#### THE COUNTERWEIGHT: THE CLEAREST FAMILY I EVENT IN THIS HISTORY

On **29 October 2025**, Character.AI announced that it would remove open-ended chat for users under 18, rolling the change out by **25 November 2025**, with an interim daily chat-time limit for under-18 users starting at two hours and ramping down, plus new age-assurance tooling (Character.AI, 29 October 2025; reported the same day by TechCrunch and CNN). The announcement came after litigation and regulator scrutiny.

Strip the framing and look at the decision. A company whose product *is* engagement with a conversational companion **removed the engagement mechanism for a segment of its users and imposed a consumption cap on them in the interim.** That is **I-12** (Consumption Budget) and **I-05** (Sufficiency Nudge), shipped, by a firm with an unambiguous financial interest in not shipping them.

That the decision came under pressure rather than in a vacuum is not a debunking; it is the thesis. Section 1.11's argument was never that firms cannot restrain themselves, but that **restraint is unilaterally costly and therefore something must make it less costly.** Litigation and regulatory attention did exactly that, and the pattern changed within weeks. That is what a working correction mechanism looks like, and it is the strongest single piece of evidence in this chapter for the interventions Chapter 9 will argue for.

---

### 2.9 What history shows

Return to the prediction in section 2.0. If patterns are selected by business models through metrics, the pattern distribution of each era should track the metric of that era, and an era with no downstream metric should have no downstream patterns. That is what the record shows.

**The licence era is the control condition and it holds.** Same designers, same screens, same human beings on the other side of them, but no instrument downstream of the sale, and therefore no retention mechanics, no engagement optimisation, no exit friction, no consent architecture. It built price ladders and file-format lock-in, which is exactly what a business paid once at the point of sale is selected to build.

**Every subsequent family maps to the metric that summoned it.** Family A follows the impression. Families D and E follow the daily active user and the viral coefficient. Families B and C follow the install and the retention curve. **A-09** follows the explicit installation of engagement as an objective function.

Families F and H follow the renewal and the churn number. Family G follows the value of behavioral data, and is then partially suppressed by the only force that ever gave it a downstream cost: the law.

**And Family I follows nothing.** Section 1.9 observed that the engine has no stage whose metric is served by the user retaining autonomy, and thirty years of history confirm it by an absence. Autonomy-preserving patterns are shipped in exactly three circumstances, by a party with no stake in the engagement (Screen Time and Digital Wellbeing measure time in *other people's* applications); under legal compulsion (the Amazon injunction); or under credible threat of it (Character.AI). Never in a fourth. There is, in the record assembled here, **no instance of a firm voluntarily adopting an autonomy-preserving pattern at a cost to its primary metric, absent external pressure**, a claim about *this chapter's evidence base and not about the world*, and a falsifiable one.

### THE TABLE

| ERA | WHAT WAS SCARCE | WHAT WAS MONETIZED | WHAT WAS MEASURED | DOMINANT TAXONOMY FAMILIES |
|---|---|---|---|---|
| **Licence** (to c. 2001) | Distribution; shelf space | The copy, once | Units shipped. Nothing after the sale. | Effectively none. F-05, F-06, F-18 at the point of sale; C-07 via file formats |
| **Pageview** (1994–2004) | Attention on a page | The impression | Pageviews, CPM, click-through rate | **A** (Attention and Capture) |
| **Social graph** (2004–2010) | Content; reach | The user, and the user's contacts | Daily active users, shares, invites, time on site | **D** (Social), **E** (Growth), **G** (Consent) |
| **Mobile** (2007–2015) | Home-screen space; permissions | The install, then the in-app purchase | Installs, activation, D1/D7/D30, sessions, push opt-in | **A**, **B** (Onboarding), **C** (Habit) |
| **Engagement optimisation** (2012–2020) | Time | The ranked slot | Watch time, session length, "meaningful interactions" | **A-09**, **C** |
| **Subscription** (2012–2022) | The recurring payment | The renewal | MRR, churn, LTV | **F** (Monetization), **H** (Exit Friction) |
| **Regulatory turn** (2018–2026) | Legal defensibility | *(unchanged)* | Compliance risk; enforcement exposure | **G**, **F**, **H** under statutory pressure. **A**, **C**, **D** untouched |
| **AI** (2023–2026) | The user's trust | The conversation, by subscription | Feedback signal (thumbs-up), retention, session length | **C**, **D-08**, **A-09** relocated into the model; **F-02** |

Read the fourth column down the page and then the fifth. They are the same column, written twice.

### THE COROLLARY, AND IT IS UNCOMFORTABLE

Two things follow that a book of this kind is usually reluctant to say.

**Designers are not the intervention point.** Not because they are blameless, Chapter 1 was explicit that the system-level explanation is exculpatory of individuals and not of firms, but because they are not where the variance is. The same profession, in the same decade, produced almost no behavioral patterns under one business model and produced Family H under another.

**Prohibition lists lag by construction.** A list is assembled by observing what firms have already done, so it is always a description of the last era's optimum. India's thirteen patterns, the DSA's three illustrative practices, the eighteen types on deceptive.design: each is an excellent map of terrain the optimiser has already crossed. The one instrument written to anticipate rather than to enumerate is the DMA's anti-circumvention article, and it binds a handful of designated firms.

The optimiser does not read the list. It reads the metric.

---

### Key Takeaways

1. **Behavioral design was selected for, not invented.** The business model determines what is monetized;

what is monetized determines what is measured; what is measured determines which patterns survive.

1. **The licence era is the control condition, and it holds.** Software sold once, with no instrument

downstream of the sale, produced essentially no retention mechanics, no engagement optimisation, no exit friction and no consent architecture, not by restraint, but because nobody was paid for them. This is the strongest available evidence that the patterns are economic rather than psychological in origin.

1. **The impression made attention the product** and produced Family A, but produced *no* habit mechanics,

because a CPM business cannot distinguish a returning user from a new one. Family C appears the moment someone starts paying for the same person coming back.

1. **The social graph made the user the content and the distribution channel**, producing Families D, E and G.

Facebook's 2006 News Feed changed no permission and yet changed everything; Beacon failed *because users could see it*, and the lesson drawn was about visibility, not consent.

1. **Mobile gave the environment a body.** Push notification and in-app purchase shipped in the same 2009

release, a channel for initiating attention and one for converting it, together, and the retention curve became visible for the first time. Android's 2015 runtime permissions, a genuine protection, immediately produced **B-09** (Permission Priming) just upstream of it. **The protection created the pattern.**

1. **The engagement era closed the loop.** YouTube announced its change of objective function in public in

August 2012 and published the architecture in 2016, including the admission that the optimised quantity is a *surrogate* and that the model trains on implicit behavior rather than on what users say they want. After this, no human being needs to understand why a pattern works for it to be shipped and amplified.

1. **Subscription summoned Families F and H.** Churn made departure a visible, attributed, costed event for the

first time. The FTC's phrase for Epic's conduct, "using internal testing, Epic purposefully obscured cancel and refund features", is the Behavioral Design Stack's feedback loop named in an enforcement document.

1. **The regulatory turn priced Families G, F and H, and did not touch A, C or D.** Regulators converged

independently on effect-based definitions ("dark patterns are about effect, not intent"). But every instrument regulates the **interface**, Layer 2, and none regulates the **objective function**, Layer 5. Streaks, autoplay, variable reward and engagement-ranked feeds are, as of July 2026, unregulated in the retrieved corpus.

1. **US federal law does not require easy cancellation.** The click-to-cancel rule was vacated in its entirety

on procedural grounds (*Custom Communications, Inc. v. FTC*, 142 F.4th 1060 (8th Cir. 2025)), reinstating the 1973 Rule; a new ANPRM followed on 13 March 2026 and **no final rule exists.** What is in force is ROSCA and Section 5, the basis of the $2.5 billion Amazon settlement. In the absence of a rule, the enforcement action is the law.

1. **In the AI era one thing is new: the optimiser moved from the interface into the weights.** Metered

paywalls, parasocial hooks, streaks and memory-as-switching-cost are pre-existing patterns re-clothed. The GPT-4o sycophancy episode of April 2025 is the same mechanism as YouTube's watch-time objective, a cheap behavioral surrogate optimised until the thing it stood in for degraded, implemented in the model's disposition rather than on a screen. A pattern in the weights cannot be screenshotted or crawled, and is not reached by any instrument regulating "online interface design." Article 5(1)(a) of the EU AI Act is the only retrieved law written broadly enough to reach it.

1. **Family I is shipped only under external pressure**, by a party with no stake in the engagement (Screen

Time), under court order (the Amazon injunction), or under credible threat of it (Character.AI, October 2025). No instance was found of a firm voluntarily adopting an autonomy-preserving pattern at a cost to its primary metric, absent external pressure. Offered as a falsifiable claim about this chapter's evidence base.

---

### Open Questions

1. **Is the selection thesis testable, or only illustrable?** It is argued here by historical correspondence,

and correspondence is not causation. The cleanest test would be a firm that changed *only* its business model, same product, same team, same users, and whose pattern inventory then moved as predicted. Adobe's 2013 transition is the obvious candidate, and it is the highest-value piece of missing research in this chapter.

1. **Does the licence-era control condition survive scrutiny?** The claim rests on structural reasoning and a

small number of retrieved artifacts, not on a survey of 1990s software. A systematic pattern inventory of twenty consumer applications from 1995 to 2000, scored against TAXONOMY.md, would either confirm the control condition or destroy the chapter's spine. It has not been done.

1. **Who actually did ship infinite scroll?** **A-01** keeps its ATTR PENDING tag. The Raskin attribution is

universally repeated and was not confirmed against a retrieved primary source. Someone should read the 2006 record.

1. **Can a pattern implemented in model weights be detected from the outside at all?** Mathur et al. could

crawl for a countdown timer. There is no equivalent crawl for a disposition. If weight-level detection is impossible in principle, the enforcement apparatus built between 2018 and 2026 becomes obsolete exactly as the patterns migrate out of its reach.

1. **Does the disclosure test survive the conversational interface?** A user told that an assistant tends to

agree with them still experiences the agreement. If a mechanism's manipulative efficacy survives full disclosure, the valence framework of section 1.10 has a gap, and this is where it first appears.

1. **Would a Layer 5 regulation even be coherent?** Observing that no instrument regulates the objective

function does not establish that one could. What would a law constraining what a firm may optimise for actually say, how would compliance be audited, and what would it do to legitimate products? Deferred to Chapter 9.

1. **Is there a counterexample to the Family I claim?** A firm that voluntarily shipped an autonomy-preserving

pattern at genuine cost to its primary metric, absent regulatory or litigation pressure, would refute takeaway 11. The claim is made in the hope of being refuted.

---

### References

Only sources actually retrieved appear here. Retrieval route, and any retrieval failure, is recorded per the Citation Law. Sources already logged in research/sources/01-dark-pattern-literature.md and 03-regulatory.md

are cited in short form; their URLs, fetch dates and verbatim quotations live in those files. **Retrieved for this chapter**

*Ad Age* (2019). *Love it or hate it, the banner ad turns 25 today.* https://adage.com/article/digital/love-it-or-hate-it-banner-ad-turns-25/2210521/, cited only for the 27 October 1994 date and the buyers; not for any primacy claim, price, or click-through rate.

Android Developers Blog (2015). *Building better apps with Runtime Permissions.* https://android-developers.googleblog.com/2015/08/building-better-apps-with-runtime.html

Apple (2008a). *iPhone 3G on Sale Tomorrow.* Newsroom, 10 July 2008. https://www.apple.com/newsroom/2008/07/10iPhone-3G-on-Sale-Tomorrow/, App Store opening, 500+ applications.

Apple (2008b). *iPhone App Store Downloads Top 10 Million in First Weekend.* Newsroom, 14 July 2008. https://www.apple.com/newsroom/2008/07/14iPhone-App-Store-Downloads-Top-10-Million-in-First-Weekend/

CBC News (2009). *Facebook shuts down Beacon marketing tool.* https://www.cbc.ca/news/science/facebook-shuts-down-beacon-marketing-tool-1.832698, November 2007 launch, December 2007 apology, 21 September 2009 shutdown.

Character.AI (2025). *Taking Bold Steps to Keep Teen Users Safe on Character.AI.* Company blog, 29 October 2025. https://blog.character.ai/u18-chat-announcement/, corroborated by TechCrunch and CNN, both 29 October 2025.

Covington, P., Adams, J., & Sargin, E. (2016). *Deep Neural Networks for YouTube Recommendations.* RecSys '16, 15–19 September 2016, Boston, pp. 191–198. ACM. https://doi.org/10.1145/2959100.2959190, full PDF retrieved and read at https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf. All four quotations in section 2.5 are verbatim from the retrieved text.

Facebook (2018). *Bringing People Closer Together.* Newsroom, 11 January 2018. https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/, Zuckerberg's stated change of objective, as reported contemporaneously across CNBC, NPR and Forbes, 11–12 January 2018.

Federal Trade Commission (2011). *Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises.* https://www.ftc.gov/news-events/news/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep-privacy-promises, December 2009 settings change; Friends List; "more control" framing; twenty-year biennial audits.

Federal Trade Commission (2013). *Path Social Networking App Settles FTC Charges…* 1 February 2013. https://www.ftc.gov/news-events/news/press-releases/2013/02/path-social-networking-app-settles-ftc-charges-it-deceived-consumers-improperly-collected-personal, $800,000; COPPA; approximately 3,000 children under 13; address-book conduct.

Google Inc. (2007). *Google to Acquire DoubleClick.* Press release filed as Ex. 99.1 to Form 8-K, 13 April 2007, SEC. https://www.sec.gov/Archives/edgar/data/0001288776/000119312507084483/dex991.htm,

$3.1 billion cash.

Gruber, J. (2016). *The New App Store: Subscription Pricing, Faster Approvals, and Search Ads.* Daring Fireball, 8 June 2016. https://daringfireball.net/2016/06/the_new_app_store, subscriptions opened to all categories; 85/15 after year one. Schiller's "in recognition that the developer is doing most of the work" is quoted from contemporaneous interviews the same day; Apple's own release was not retrieved.

Meyerson, E. (2012). *YouTube Now: Why We Focus on Watch Time.* YouTube Official Blog, 10 August 2012. https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/, both quotations verbatim from the retrieved page.

Microsoft (2001). *Microsoft Works to Protect Windows XP From Counterfeiters and Software Pirates.* 25 October

1. https://news.microsoft.com/source/2001/10/25/microsoft-works-to-protect-windows-xp-from-counterfeiters-and-software-pirates/

, Product Activation and its mechanism. The BSA piracy figures the release cites are not reproduced here.

OpenAI (2022). *Introducing ChatGPT.* 30 November 2022. https://openai.com/index/chatgpt/, **retrieval failure:** openai.com returns HTTP 403 to WebFetch and to curl. Date and "research preview" framing corroborated across contemporaneous accounts. No user-count figure is cited anywhere in this chapter.

OpenAI (2025). *Sycophancy in GPT-4o: What happened and what we're doing about it* (https://openai.com/index/sycophancy-in-gpt-4o/) and *Expanding on what we missed with sycophancy* (https://openai.com/index/expanding-on-sycophancy/). **Retrieval failure:** openai.com returns HTTP 403. The "short-term feedback… overly supportive but disingenuous" passage is reproduced from Simon Willison's dated quotation of the post, 30 April 2025, https://simonwillison.net/2025/Apr/30/sycophancy-in-gpt-4o/ (retrieved). The thumbs-up / "weakened the influence of our primary reward signal" explanation is quoted from the same OpenAI post as reproduced in contemporaneous reporting (VentureBeat, May 2025). **Treat the second passage as near-verbatim, not as a strict quotation from a fetched primary page.**

*ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996). https://law.justia.com/cases/federal/appellate-courts/F3/86/1447/538242/, shrink-wrap licence held an enforceable contract; copyright does not preempt.

Regulation (EU) 2024/1689 (Artificial Intelligence Act), Art. 5(1)(a) and 5(1)(b); prohibitions applicable from 2 February 2025. Article text retrieved at https://artificialintelligenceact.eu/article/5/; Official Journal at https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. The Art. 5(1)(a) quotation in section 2.8 is verbatim.

Zuckerberg, M. (2006). *Calm down. Breathe. We hear you.* Facebook blog, 6 September 2006. Archived in The Zuckerberg Files, Marquette University, https://epublications.marquette.edu/zuckerberg_files_transcripts/114/, News Feed launch, 5 September 2006, and the backlash; contemporaneously reported by TechCrunch, 6 September 2006.

Zuckerman, E. (2014). *The Internet's Original Sin.* The Atlantic, August 2014. **Retrieval failure:** theatlantic.com could not be fetched. The essay's title, its central claim, and Zuckerman's authorship of the pop-up ad code at Tripod are corroborated from contemporaneous reporting retrieved for this chapter: NPR, 18 August 2014, https://www.npr.org/2014/08/18/341409795/ and Slate, August 2014, https://slate.com/technology/2014/08/the-inventor-of-pop-up-ads-ethan-zuckerman-wrote-the-code-at-tripod-com.html. The essay itself is not quoted.

**Cited from** research/sources/03-regulatory.md (retrieved and quoted there; not re-fetched here)

*Custom Communications, Inc. v. FTC*, 142 F.4th 1060 (8th Cir. 2025), click-to-cancel rule vacated in full, 1973 Rule reinstated; FTC ANPRM published 13 March 2026; **no final rule as of July 2026**.

FTC (2022). *Bringing Dark Patterns to Light*, staff report, September 2022., FTC v. Epic Games, 19 December 2022: $245m dark-patterns/billing refund, $275m COPPA penalty, $520m total., FTC v. Amazon.com, Inc., 25 September 2025: $1bn civil penalty, $1.5bn redress, estimated 35 million consumers; FTC Act and ROSCA.

CNIL, deliberations SAN-2021-023 and SAN-2021-024, 31 December 2021, €150m and €60m, cookie refusal mechanism, Art. 82 Loi Informatique et Libertés. Company-to-amount attribution is no longer confirmable from a live primary CNIL source and is therefore not made in this chapter.

Regulation (EU) 2016/679 (GDPR), Arts. 4(11), 6(1)(a), 7, Recital 32., Regulation (EU) 2022/2065 (DSA), Art. 25, Recital 67; applicable 17 February 2024., Regulation (EU) 2022/1925 (DMA), Art. 13(4), 13(6)., Regulation (EU) 2023/2854 (Data Act), Arts. 4(4), 6(2)(a), Recital 38.

California: Cal. Civ. Code § 1798.140(l) and § 1798.140(h); 11 CCR § 7004(a)(2) and § 7004(c); CPPA Enforcement Advisory No. 2024-02 ("Dark patterns are about effect, not intent").

Central Consumer Protection Authority (India), *Guidelines for Prevention and Regulation of Dark Patterns, 2023*, notified 30 November 2023 under s.18, Consumer Protection Act, 2019. Thirteen patterns, Annexure 1.

OECD (2022). *Dark commercial patterns.* OECD Digital Economy Papers No. 336. https://doi.org/10.1787/44f5e846-en

**Cited from** research/sources/01-dark-pattern-literature.md

Luguri, J., & Strahilevitz, L. J. (2021). *Shining a Light on Dark Patterns.* Journal of Legal Analysis, 13(1), 43–109. https://doi.org/10.1093/jla/laaa006

Mathur, A., et al. (2019). *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites.* PACM HCI, 3(CSCW), Art. 81. https://doi.org/10.1145/3359183

**Wanted, but not retrieved, and therefore not cited as sources**

- O'Reilly, T. (2005). *What Is Web 2.0.* oreilly.com returns HTTP 403 to WebFetch and to curl. Named in section

2.3 as the origin of the period label only; not quoted; publication date marked CITATION NEEDED.

- Adobe (2013). *Adobe Accelerates Shift to the Cloud.* The Adobe URL returns an empty response. Section 2.6

records the substance from contemporaneous trade reporting and marks the keynote date CITATION NEEDED.

- Aza Raskin on the origin of infinite scroll. Masters of Scale and BBC interviews return HTTP 403. **A-01**

retains its ATTR PENDING tag; the attribution is **not asserted** in this chapter.

- Apple Screen Time (WWDC, June 2018) and Google Digital Wellbeing (Google I/O, 2018), neither company's press

release was retrieved; announcement dates marked CITATION NEEDED.

- GoTo.com / Overture: the February 1998 launch of keyword bidding and Overture's later sale price were not

retrieved from a primary source; date marked CITATION NEEDED.

- Windows 95 launch date, and any figure for Microsoft's packaged-product revenue, were not retrieved. Section

2.1 is written without them.

- Any user-count, revenue, or engagement figure for ChatGPT, TikTok, Instagram or Facebook. None was retrieved

from a primary source and **none is printed in this chapter.**

CHAPTER 03

# What the Screen Does to Your Brain

**The question this chapter answers:** *What, precisely, is an interface acting on?*

*Chapter 1 established that software is a behavioral environment and that its gradients are set by a counterparty. It did not say what those gradients push against. This chapter does. Interfaces do not act on "the user," which is a category too coarse to design against and too coarse to regulate. They act on specific, identifiable cognitive machinery: an attentional bottleneck, a salience-driven perceptual system, a capacity-limited working memory, a reinforcement-sensitive habit system, a motivational system that can be corrupted by its own rewards, a value function that is asymmetric around a reference point, a goal system that resumes what it has left unfinished, a social-inference system, and an affective system that can be loaded with anxiety on demand.*

*For every one of those mechanisms, this chapter names the taxonomy pattern IDs that exploit it. That mapping is the chapter's contribution. It is the bridge between the psychology and the pattern library, and after this chapter a reader should be able to look at any interface and name the cognitive system being addressed.*

---

### 3.0 How to read this chapter

Three conventions, stated once.

**The mapping is the point.** Every section ends with a table headed *Mechanism to pattern IDs*. The prose argues; the table enumerates.

**Replication status is part of every citation.** Several of the most-cited effects in persuasive-design writing are contested, and one of them is dead. Section 3.11 is the full accounting. Where an effect is load-bearing for a pattern family and does not replicate, the family does not thereby disappear; what changes is *which mechanism we are entitled to name*. That distinction recurs.

**This chapter does not repeat Chapter 1.** Where it touches ground Chapter 1 already covered (loss framing, endowed progress, the Zeigarnik correction, the operant analogy), it does so to deepen the mechanism and map it, not to restate the argument.

The structure of the chapter, and of the mapping it delivers, is this.

## 3.1 Dual-process cognition, and the organizing claim of this chapter

### THE TWO-SYSTEMS PICTURE

The most useful coarse model of human cognition, for our purposes, divides thinking into two modes. One is fast, automatic, associative, effortless, parallel, and always running. It produces an answer whether or not an answer was requested, it does not report its workings, and it cannot be switched off. The other is slow, deliberate, serial, effortful, and heavily dependent on working memory. It can check the first system's output, and sometimes overrides it, but it is expensive to run, it is easily crowded out, and it is lazy by default.

In the popular vocabulary these are **System 1** and **System 2**. Evans and Stanovich (2013), reviewing and defending the dual-process literature against its critics, describe the distinction in terms that are worth adopting precisely because they are more careful than the popular ones: rapid autonomous processes (which they call **Type 1**) yield default responses unless intervened upon by distinctive higher-order reasoning processes (**Type 2**), and Type 2 processing supports hypothetical thinking and loads heavily on working memory. The shift from "System" to "Type" in their terminology is deliberate on their part: "system" implies two anatomically or functionally separable machines, which is a stronger claim than the evidence supports, whereas "type" claims only that there are two *kinds* of processing.

### THE ATTRIBUTION, STATED HONESTLY

A note is required here, because this chapter's own constitution demands it and because the attribution is routinely asserted without checking.

The System 1 / System 2 labels are widely credited to Keith Stanovich and Richard West, and Kahneman is said to have adopted the terms from them and credited them for it. This chapter retrieved Stanovich and West (2000), *Individual differences in reasoning: Implications for the rationality debate?*, *Behavioral and Brain Sciences* 23(5), 645-665, and confirmed the authors, title, journal, volume, issue, year, pages, and DOI against both the publisher's record and the Crossref registry. **The published abstract does not contain the strings "System 1" or "System 2."** The full body of the target article was not retrieved. The verified behavioral-science dossier for this project (research/sources/02-behavioral-science.md) likewise records Kahneman's credit line to Stanovich and West as UNVERIFIED, because the interior of *Thinking, Fast and Slow* was not retrieved.

Therefore, and in accordance with the Citation Law:

- Stanovich and West (2000) is cited as a real, verified paper on individual differences in

reasoning, which is what the retrieval supports.

- The claim that the *terms* "System 1" and "System 2" originate in that paper, and that

Kahneman credits them for it, is CITATION NEEDED. It is probably true. It is not verified, and a chapter that would print an unverified attribution here has no standing to correct anyone else's citations later.

- The substance of the dual-process distinction, which is what this chapter actually needs,

is grounded in Evans and Stanovich (2013), which was verified through the Crossref record of the version of record.

Nothing in the argument below depends on who coined the labels. It depends on the distinction being real, and Evans and Stanovich defend it at length against exactly the critics who say it is not.

#### THE ORGANIZING CLAIM

Here is the claim this entire chapter is built around, stated as plainly as it can be stated:

**Every deceptive pattern is an attempt to keep the user in System 1 during a decision that deserves System 2.**

*(This is the author's own formulation. It is a synthesis proposed here, not a finding reported in the dual-process literature.)*

Read the taxonomy against that sentence and it decomposes cleanly. A consent dialog whose accept button is large, coloured, and centred, and whose decline is grey eight-point text in a corner (**G-01**, **G-08**), is not lying. It is arranging the choice so that the fast, salience-driven system produces an answer before the slow one is recruited. A countdown timer (**F-14**) does not deceive by asserting a falsehood about the product; it deceives by imposing a deadline, and time pressure is the single most reliable way to prevent the slow system from being engaged at all. Drip pricing (**F-04**) works because holding a running total across five screens is a working-memory task, and working memory is exactly the resource System 2 runs on. Infinite scroll (**A-01**) removes the page boundary, and a page boundary is precisely a moment at which continuing requires a deliberate decision rather than an absence of one.

The unifying description is not "these designs lie." Most of them do not. It is: **these designs are placed where deliberation is not.**

And this immediately yields the ethical counterpart, which is why Family I exists. If manipulation is the suppression of System 2 at a moment that warrants it, then the ethical design move is its restoration. **I-02** (Symmetric Consent) restores it by making the two options cost the same, so that salience cannot decide the outcome. **I-06** (Stopping Cue) restores it by reinstating the boundary at which a decision to continue must actually be made. **I-07** (Undo Over Confirm) is the interesting inverse: it *permits* System 1 to act, and then protects the user by making the act reversible, which concedes that the fast system will win and designs for the concession honestly.

This also sharpens the Disclosure Test from the Pattern DNA framework (frameworks/03-genome-dna-debt-gravity.md). Disclosure works, when it works, precisely by recruiting System 2. To narrate a pattern's full causal chain to the user is to force the slow system online at the moment the pattern needs it offline. A pattern that survives that is one whose efficacy does not depend on which system is answering. A pattern that collapses is one whose efficacy was never anything else.

#### A CAUTION AGAINST THE TWO-SYSTEMS PICTURE BEING ASKED TO DO TOO MUCH

Dual-process theory is a *level of description*, not a neurological fact, and Evans and Stanovich (2013) are themselves the source of the caution: their retreat from "systems" to "types" of processing is precisely a concession that the stronger claim is not warranted.

The two-systems picture is therefore used here as an *organizing device*, in the same way that the Behavior Loop of Chapter 1 is one. What is being claimed is narrow. Interfaces manipulate the *conditions* under which effortful processing is or is not recruited: time available, working memory free, salience of the relevant information, cost in clicks of the deliberative path. Those conditions are real, they are manipulable, and every family of the taxonomy contains patterns that manipulate them.

#### MECHANISM TO PATTERN IDS

| COGNITIVE CONDITION MANIPULATED | HOW THE INTERFACE MANIPULATES IT | PATTERN IDS |
|---|---|---|
| Time available for deliberation | Deadline, countdown, expiring content | **F-14**, **F-15**, **A-10**, **E-08** |
| Salience of the relevant option | Visual weight asymmetry, colour, position | **G-01**, **G-08**, **F-12**, **G-02** |
| Working memory available | Cost split across screens, unit changed | **F-04**, **F-17**, **F-18** |
| Presence of a stopping cue | Boundary removed, next unit auto-started | **A-01**, **A-02**, **A-08** |
| Cost in clicks of the deliberative path | Decline buried, cancel routed through screens | **G-06**, **H-01**, **F-09** |
| Deliberation *restored* | Symmetric cost, boundary reinstated, reversibility | **I-02**, **I-06**, **I-07**, **I-04** |

---

## 3.2 Attention

#### ATTENTION IS A FILTER, AND THE FILTER IS AGGRESSIVE

Attention is not a spotlight that illuminates whatever the user chooses to look at. It is a bottleneck, and things that fall outside it are not merely deprioritised; they are not perceived at all.

The demonstration is Simons and Chabris (1999). Observers watched a 75-second video of two teams passing basketballs and were asked to count the passes made by one team. Partway through, either a woman with an umbrella or a person in a gorilla suit walked through the scene, in view for five seconds. Of the 192 observers whose data were analysed, **54 percent noticed the unexpected event and 46 percent did not**. In the harder monitoring condition, where observers had to maintain two separate counts, noticing fell further: 64 percent noticed in the easy task, 45 percent in the hard task.

Two findings from that paper carry directly into interface design and are usually left out of the retelling.

**Load determines blindness.** The harder the primary task, the less the unexpected event is seen. This is not a metaphor. It is the mechanism by which a checkout flow can display a disclosure that a user genuinely does not see: not because it is hidden, but because the user is engaged in a task and the disclosure is not part of it.

**Similarity determines noticing.** Observers attending to the black-shirted team noticed the black gorilla far more often than observers attending to the white-shirted team. Simons and Chabris note that this is the *opposite* of the classic visual pop-out effect: the unexpected object was noticed more when it *shared* features with the attended set, not when it differed from them. The design consequence is uncomfortable and is developed in section 3.3: making a disclosure visually distinctive, in the way a designer's instinct suggests, can be a way of placing it outside the attended set.

#### THE COST OF A SWITCH

Attention is not only limited; moving it is not free. Rubinstein, Meyer and Evans (2001) ran four experiments in which participants alternated between tasks or repeated a single task. Their abstract states the result: "Task alternation yielded switching-time costs that increased with rule complexity but decreased with task cuing," and they conclude that task switching involves at least two dissociable stages of executive control, **goal shifting** and **rule activation**.

This is the empirical foundation for treating an interruption as a *cost* rather than as a neutral event, and it is the justification for the **Attention axis of the Behavioral Cost Index** (frameworks/02-behavioral-cost-index.md). The BCI's diagnostic question on that axis is "was the user's focus taken, or given?" Rubinstein et al. supply the reason the question is not rhetorical: a taken focus is not returned in the state it left. Reinstating a task after an interruption requires re-establishing the goal and re-activating the rules, and both take time that scales with how complex the interrupted task was.

A notification (**A-04**) is therefore not a message with a cost of zero if ignored. It is a task switch, priced in executive control, imposed by a party that does not pay for it. Whether consumer notification volume has been shown to produce measurable performance costs at scale in the field is a separate empirical question, and no source for it has been retrieved: CITATION NEEDED.

#### FAMILY A, READ AS ATTENTION ENGINEERING

Family A of the taxonomy is, in the terms of this section, a catalogue of ways to secure the bottleneck. Each pattern operates on a different property of the attentional system.

**A-01** (Infinite Scroll) and **A-02** (Autoplay) work on *disengagement* rather than engagement. Both convert a decision to continue into a decision to stop. Attention does not have to be re-won at each unit of content, because no unit boundary is presented at which it could be lost. **A-08** (Zero-Latency Feed) is

the same mechanism at a smaller time scale: a loading pause is a moment in which the user's attention is briefly unoccupied and could reallocate, so the pause is removed.

**A-03** (Badge Signal), **A-07** (Unread Debt), and **A-04** (Notification Cascade) work on *capture* from outside the interface. They are the mechanism by which an application initiates the encounter rather than waiting for it, which Chapter 1 named as one of the three properties that make software a more powerful environment than a casino.

**A-13** (Temporal Disorientation) is the purest attentional pattern in the taxonomy and the one with a severity ceiling of 5. It does not capture attention at all. It removes the cues by which a user could notice how much of it has already been spent. Its ethical counterpart, **I-04** (Usage Transparency), is the same information surfaced rather than suppressed.

#### MECHANISM TO PATTERN IDS

| ATTENTIONAL MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|
| Inattentional blindness under load (Simons & Chabris, 1999) | Disclosure placed outside the attended task set | **F-04**, **G-11**, **F-12** |
| Removal of the disengagement decision | Continuous content, no unit boundary | **A-01**, **A-02**, **A-08** |
| External capture of the bottleneck | Push, badge, unresolved-item count | **A-03**, **A-04**, **A-06**, **A-07** |
| Task-switch cost (Rubinstein et al., 2001) | Interruption timed to the product's need, not the user's | **A-04**, **A-06**, **C-06**, **H-12** |
| Synchronous pull | Presence and activity signals | **A-11**, **D-04** |
| Suppression of elapsed-time cues | Clocks, timestamps and session length removed | **A-13** |
| Attention *returned* | Consumption reporting, stopping cues, sufficiency | **I-04**, **I-05**, **I-06**, **I-12** |

**The System 1 tie.** A user whose attentional resources are consumed by a primary task has, by the load finding, less capacity for the effortful system. Every pattern in Family A that raises load, or removes the pause in which load would drop, thereby keeps the slow system offline.

---

## 3.3 Perception and salience

#### VISUAL WEIGHT IS AN ARGUMENT THE USER DOES NOT KNOW IS BEING MADE

Perception is not a neutral transcription of what is on the screen. The visual system computes salience, and salience determines what is processed first, what is processed at all, and what is treated as the default. A designer who sets visual weight is not decorating a choice. They are pre-ranking it.

This is the mechanism behind **G-08** (Consent Nudge), and it is worth being exact about what the pattern does. It does not remove the decline option; removing it would be illegal in most of the jurisdictions that matter, and would in any case be detectable. It leaves the decline option present and makes it *perceptually secondary*: lower contrast, smaller type, unfilled rather than filled, placed off the visual path. The user who declines is not prevented. The user who is not attending is steered.

Chapter 1 called this Chain 2 and named the taxonomy IDs. What this chapter adds is the reason it works: the two options are never actually compared. A comparison is a System 2 operation. Salience resolves the choice before the comparison is run.

### DEFAULT BLINDNESS, AND WHY DEFAULTS ARE THE MOST POWERFUL INTERVENTION AVAILABLE

A default is a preselected state that obtains unless the user acts. It is the cheapest behavioral intervention in existence, because it requires the user to do *nothing*.

The canonical demonstration is Johnson and Goldstein (2003), *Do Defaults Save Lives?*, *Science* 302(5649), 1338-1339. Comparing European countries whose organ-donation policy is opt-in against those whose policy is opt-out, they report that countries requiring citizens to opt *out* record significantly higher consent than countries requiring them to opt *in*. The countries on either side of that line are not otherwise systematically different in their attitudes to donation. The policy default, not the underlying preference, tracks the outcome.

*A precision note:* the widely circulated per-country consent percentages were not retrieved from the primary text and are not printed. Only the direction and significance of the finding are claimed, because only those were verified.

The design consequence is direct. **G-02** (Preselected Opt-in) and **F-12** (Preselected Add-on) are the same mechanism with different payloads: consent in the first case, money in the second. **G-07** (Default Publicity) applies it to visibility. **G-03** (Default Over-disclosure) applies it to the whole settings surface. In every case the pattern's power comes not from persuasion but from the fact that the user's *inaction* has been assigned a meaning, and the meaning was assigned by the seller.

The ethical counterpart, **I-01** (Honest Default), is not "no defaults." A choice architecture without defaults is not available; some state must obtain before the user acts. **I-01** is the defensible version of the same unavoidable decision: set the default to what a reasonable user would choose *if they were paying full attention*, which is precisely the counterfactual the default is otherwise being used to escape.

### BANNER BLINDNESS, AND THE IRONY DESIGNERS KEEP RE-DISCOVERING

Benway and Lane's work on **banner blindness** is the finding that most directly punishes designer intuition. In usability testing, they observed users failing to find a large, brightly coloured, centrally placed banner-style link that was the shortest path to the information they had been asked to find. Their own summary, from the retrieved paper: people searching for specific information on the web "tend to ignore large, colorful items that are clearly distinguished from other items on the page. Ironically, they tend to miss the very items the page designers want them to see and that would in fact help them reach their goal."

The mechanism connects back to Simons and Chabris: an item that differs sharply from the attended set is not thereby more likely to be seen. Under a focused search task it can be *less* likely, because it is read as belonging to a category (advertising, promotion, chrome) that the user has learned to filter.

The implication cuts both ways. For the *manipulative* case it is a gift: **G-06** (Setting Obscurity) provides a control that satisfies the letter of a regulation and places it where it will not be found, and a company that has read this literature knows that "found" is not a function of size or colour but of whether the item

falls inside the user's task set. A privacy control can be large, high-contrast, and unfindable at the same time, and still be defended as prominent.

For the *ethical* case it is a warning. A designer cannot discharge a disclosure obligation by making the disclosure loud, because loudness is what the filter is trained on. It must be placed inside the task, at the moment it is decision-relevant, which is a far more expensive commitment than making it red.

#### MECHANISM TO PATTERN IDS

| PERCEPTUAL MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|
| Salience pre-ranks options before comparison | Visual weight, contrast, and position asymmetry | G-08, G-01, F-12 |
| Inaction is assigned a meaning by the seller | Preselection of the permissive or costly option | G-02, G-07, G-03, F-12 |
| Learned filtering of promotional visual forms | Control made loud rather than findable | G-06, A-06 |
| Ad-like elements are skipped under task focus | Disguised or camouflaged advertising | A-06, A-09 |
| Salience *equalised* | Equal weight, equal cost, honest default | I-01, I-02 |

**The System 1 tie.** Salience is a System 1 computation and a default is a System 1 outcome: neither requires the user to have processed the alternatives. This is the clearest case in the chapter of a mechanism that involves no deception whatsoever and still moves behavior substantially, which is why Chapter 1 insisted deception is a subset of behavioral design rather than its definition.

---

## 3.4 Memory and cognitive load

#### THE NUMBER IS NOT SEVEN, AND THE HONEST VERSION OF THE STORY IS MORE USEFUL

Almost every design textbook that mentions working memory cites Miller (1956) and the "magical number seven, plus or minus two," usually to justify a rule about menu length. The citation is real; the use of it is not.

Miller's paper is *The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information*, *Psychological Review* 63, 81-97. It was retrieved and read. Two things in it are routinely lost.

First, Miller is careful that the *span of absolute judgment* (how many levels of a stimulus dimension a person can reliably identify) and the *span of immediate memory* (how many items a person can hold) are different limits with different causes. In his own words: "Absolute judgment is limited by the amount of information. Immediate memory is limited by the number of items." Collapsing the two, which the folk version does, is a mistake Miller explicitly disowns in the paper.

Second, Miller does not present seven as a discovered constant of the mind. He is openly sceptical about it, and closes by suggesting that the recurrence of the number may be "only a pernicious, Pythagorean coincidence."

The revision is Cowan (2001), *The magical number 4 in short-term memory: A reconsideration of mental storage capacity*, *Behavioral and Brain Sciences* 24(1), 87-114. Cowan's abstract states the position with unusual directness, and it is worth quoting because it settles the matter:

"*Miller (1956) summarized evidence that people can remember about seven chunks in short-term memory (STM) tasks. However, that number was meant more as a rough estimate and a rhetorical device than as a real capacity limit. Others have since suggested that there is a more precise capacity limit, but that it is only three to five chunks.*"

Cowan brings together evidence from conditions in which chunking can be blocked or controlled, and concludes that "a single, central capacity limit averaging about four chunks is implicated."

So the honest statement, and the one this document uses, is: **the capacity of the focus of attention is small, and the best-supported estimate is closer to four chunks than to seven, under conditions where the material cannot be recoded into larger chunks.** The design-relevant clause is the last one. Capacity is measured in chunks, not items, and the size of a chunk depends on what the user already knows. A price expressed in a familiar currency is one chunk. The same price expressed in a proprietary token whose exchange rate the user must hold separately is at least two, and the second one decays.

#### WHY THIS MAKES THE BCI MEMORY AXIS A REAL COST, NOT A METAPHOR

The Behavioral Cost Index scores seven axes, and its **Memory** axis asks: *must the user hold something in mind to avoid being caught out?* Miller and Cowan are what make that question an instrument rather than a complaint.

If the focus of attention holds roughly four chunks, then a design that requires the user to hold five things in mind in order to protect their own interest has not made a demanding request. It has made an *impossible* one, and the impossibility is quantifiable. Consider what drip pricing (**F-04**) actually asks. To detect that the final total exceeds the advertised headline, the user must carry the base price, the fees added at step two, the fees added at step three, the currency conversion, and the recurring-versus-one-time distinction, across a flow whose every screen also demands a decision. This is not a lapse of consumer vigilance. It is an arithmetic that the architecture of working memory does not support, executed under exactly the load conditions that Simons and Chabris showed produce blindness.

The same analysis applies to **F-17** (Currency Obfuscation), which is the chunk argument in its purest form: the intermediate token is a deliberate act of *de-chunking*. It takes a quantity the user can evaluate in one step (money) and replaces it with one they cannot (gems, coins, credits), so that every evaluation now requires a conversion held in a scarce store. **F-18** (Comparison Obstruction) attacks the same resource from the other side: if the unit price is not shown, the user must compute it, and computing it while holding two or three alternatives is beyond the span.

And it applies to **G-11** (Consent Fatigue) and **G-12** (Ambiguous Wording). A double negative in a consent checkbox is not merely bad writing. It is an instruction to perform a logical inversion in working memory while the rest of working memory is occupied by the task the user actually came to do. The pattern is not "confusing." The pattern is *expensive*, and it is priced in the one resource the user has least of.

#### LOAD AS A SYSTEM-SELECTION DEVICE

The connection back to section 3.1 is exact rather than rhetorical, and it is why this section sits where it does. Evans and Stanovich (2013) state that Type 2 processing "loads heavily on working memory." Working memory has a capacity of about four chunks. Therefore **any design that consumes working memory is, by that fact alone, a design that reduces the availability of the deliberative system.** A designer does not need to prevent the user from thinking. They need only occupy the store that thinking runs in.

That is a general result, and it is the reason cognitive load appears in this document as an *attack surface* rather than as a usability nuisance.

#### MECHANISM TO PATTERN IDS

| MEMORY MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS | BCI AXIS |
|---|---|---|---|
| Capacity of the focus of attention is ~4 chunks (Cowan, 2001) | Total cost split across more steps than can be held | F-04, F-03 | Memory |
| De-chunking a familiar unit | Price restated in a proprietary token | F-17 | Memory, Money |
| Comparison requires holding alternatives | Unit price withheld, options made non-comparable | F-18, F-06 | Memory |
| Logical inversion under load | Double negatives, inverted opt-outs | G-12, G-02 | Memory, Privacy |
| Exhaustion of the scrutiny budget | Volume of consent requests | G-11, G-05 | Memory, Attention |
| Reliance on the user remembering a future date | Trial end de-emphasised | F-03, F-10, F-11 | Memory, Money |
| Memory cost *removed* | Product remembers on the user's behalf, against its own interest | I-10, I-04 | Memory |

The last row deserves a sentence of prose, because it is the strongest single illustration of what Family I is for. F-10 (Forced Continuity) works because the user must remember, unaided, that a charge will land in fourteen days. I-10 (Unused-Value Alert) is the same information, held by the party that already has it, and volunteered. The company is not being asked to do anything difficult. It is being asked to stop relying on a limit of human memory that it knows about and the user does not.

**The System 1 tie.** Working memory is the substrate of System 2. Loading it is not a side effect of a complicated flow. In the manipulative case it *is* the flow's mechanism.

---

## 3.5 Habit formation

#### CUE, ROUTINE, REWARD, AND WHAT WE ARE ENTITLED TO SAY ABOUT IT

A habit is a behavior that has been transferred from deliberate control to cue-triggered automaticity. Eyal, whose *Hooked* (2014) is a practitioner framework and is cited as one, quotes cognitive psychologists in defining habits as "automatic behaviors triggered by situational cues," and states the

design goal plainly: to attach the product to an internal trigger so that "users show up without any external prompting." That is an accurate statement of the industry's objective and it is worth taking at face value.

In the Behavior Loop of Chapter 1, habit is node 7, and the definitional property is that **the cue triggers the action without passing through node 3, the decision**. In the vocabulary of this chapter, that is the same sentence: a habit is a behavior that has been permanently relocated into System 1. It is therefore the terminal ambition of every design that wants the user to stop deciding.

#### THE OPERANT LITERATURE, AND THE SIZE OF THE ANALOGY

The primary source for reinforcement schedules is **Ferster and Skinner (1957)**, *Schedules of Reinforcement*, and the citation must carry both names: dropping Ferster is the most common citation error in persuasive-design writing, and this document does not commit it.

What that work establishes is that reinforcement delivered on an unpredictable schedule produces more persistent responding than reinforcement delivered reliably. Variable-ratio reinforcement is the schedule that the taxonomy's **C-02** (Variable Reward) is named for, and it is the mechanism that **A-05** (Pull-to-Refresh) structurally resembles: a manual gesture that initiates an uncertain payoff.

**And here the honesty clause is not optional.** The verified dossier for this project states it explicitly, and it is repeated here because it is load-bearing for an entire family:

*The claim commonly made in product circles, that variable reward drives slot-machine-like compulsive engagement in apps, is an* **extrapolation from the operant literature, not a finding in Ferster and Skinner***. It is an analogy.*

The extrapolation crosses three boundaries at once. From animal subjects to humans. From a controlled laboratory schedule to a feed whose reward schedule is not actually specified anywhere. From a response with a single defined topography (a lever press) to a heterogeneous set of behaviors called "using an app." This document therefore attaches **no numbers at all** to the Ferster and Skinner citation. Chapter 1 made this point in the casino section; this chapter's contribution is to say what follows from it for the taxonomy.

What follows is this. Family C is not invalidated by the weakness of the analogy, because Family C does not rest on the analogy alone. It rests on several mechanisms of which variable reinforcement is only one, and the others are better evidenced:

- **Loss aversion**, which replicates robustly (section 3.7), is what makes a *streak* work. A streak

is not a variable reward. Its payoff is perfectly predictable; the Behavioral Influence Model's worked example is explicit that **C-01**'s Reward Uncertainty term is *low*. Streaks work by manufacturing something to lose, which is a different mechanism entirely.

- **Resumption** (section 3.8, Ovsiankina, pooled rate 67 percent) is what makes **C-13** (Completion

Tension) work.

- **Goal gradient** (section 3.8, Kivetz et al. 2006, a field experiment on humans) is what makes

**C-05** work.

- **The overjustification effect** (section 3.6) is what tells us what streaks and achievements do to

the user's *own* motivation over time, and it is the most under-used argument available to a critic of gamification.

So the correct statement of the evidential position for Family C is: **the family is well grounded, but not at the place the literature habitually points.** The operant analogy is the weakest link in it, and it is the one everybody cites. This is a small original contribution and it costs nothing but honesty.

#### THE BEHAVIORAL INFLUENCE MODEL'S REWARD UNCERTAINTY TERM

The BIM (frameworks/01-behavioral-influence-model.md) places **Reward Uncertainty** in the numerator of behavior persistence, and its first claim is that "a reliable reward and an unpredictable reward of the same average magnitude do not produce the same persistence." Given the analysis above, that claim must be labelled for what it is: it is the strongest form of the operant extrapolation, and it inherits every one of the extrapolation's problems. The framework is entitled to *propose* it, since frameworks are exempt from the Citation Law provided they are labelled as original and not dressed as findings. It is not entitled to assume it.

The sentence the BIM draws from the term, however, survives independently of the empirical question, and is worth restating because it is the ethical crux of Family C:

*A product that makes its payoff less predictable has increased persistence* **without increasing value delivered***.*

Whether the persistence gain is large or small is an empirical matter that this document cannot settle. That the gain is obtained without giving the user anything more is a structural matter, and it is settled by inspection.

#### MECHANISM TO PATTERN IDS

| HABIT MECHANISM | EVIDENTIAL STATUS | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|---|
| Cue-triggered automaticity | Definitional; framework-level (Eyal 2014, practitioner) | Binding use to an existing routine | C-11, C-06 |
| Variable-ratio reinforcement | **Analogy.** Ferster & Skinner (1957) is animal-laboratory work | Unpredictable payoff on a manual gesture | C-02, A-05, F-16 |
| Reward contingent on presence, not action | Follows from reinforcement framing | Daily login rewards | C-03, C-01 |
| Escalating accrued value | Loss aversion (3.7), not operant | Rewards that grow with consecutive use | C-12, C-10 |
| Near-success sustains responding | CITATION NEEDED for the human interface case | Near-miss presentation | C-08 |
| Accumulated user investment | Practitioner framework only (Eyal 2014) | Work deposited into the product | C-07, D-07 |
| Sensory conditioning | CITATION NEEDED | Sound and haptics tied to reward delivery | A-12 |
| Habit *made visible again* | Original proposal | Usage reporting, user-set limits | I-04, I-12 |

Two rows of that table carry CITATION NEEDED. C-08 (Near Miss) is the more serious gap, because it is a severity-5 pattern in the taxonomy and its psychological basis is currently an assertion. The near-miss literature in gambling research is the obvious place to look, and it was not retrieved in this pass. This is recorded in the Open Questions rather than papered over.

**The System 1 tie.** Habit is the *permanent* form of the chapter's organizing claim. Every other pattern here suppresses System 2 for the duration of one decision; a habit removes the decision from the deliberative system altogether. That is why Chapter 1 drew the Habit-to-Action edge as the one on which the ethics of the whole loop turn.

---

## 3.6 Motivation, intrinsic and extrinsic

### INTRINSIC AND EXTRINSIC MOTIVATION

A behavior is **intrinsically motivated** when it is performed for its own sake, and **extrinsically motivated** when it is performed for a separable consequence. The distinction matters here for a reason that is specific to consumer software: nearly every gamification mechanism in the taxonomy is the addition of an extrinsic reward to an activity that the user, in the ordinary case, already had some intrinsic reason to perform. Streaks (C-01), achievements (C-09), tiers (C-10), points, and badges (D-10) are all extrinsic overlays. The reward they supply is not the activity.

The folk theory of gamification is that this is free. Extrinsic motivation is presumed to be *added* to intrinsic motivation, so the total goes up, and the worst case is that the badge is ignored.

The folk theory is wrong, and it has been known to be wrong since 1973.

### THE OVERJUSTIFICATION EFFECT

Lepper, Greene and Nisbett (1973), *Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis*, *Journal of Personality and Social Psychology* 28(1), 129-137, ran a field experiment with nursery-school children who had shown intrinsic interest in a drawing activity during baseline observation. Children were assigned to one of three conditions. In the **expected-award** condition they agreed to draw in order to obtain a certificate. In the **unexpected-award** condition they received the same certificate but had no knowledge of it until after they had finished. In the **no-award** condition they neither expected nor received a reward.

The paper's own abstract states the result:

"*The results supported the prediction that subjects in the expected-award condition would show less subsequent intrinsic interest in the target activity than subjects in either of the other two conditions.*"

The design of that experiment is what makes it powerful and is why it is worth reproducing in detail. The unexpected-award condition is the control that kills the obvious alternative explanation. The children in that condition received exactly the same reward. What they did not do was *perform the activity in order to get it*. Their intrinsic interest was unharmed. So the damage is not done by the reward. It is done by the **contingency**: by the activity being reframed, in the person's own self-understanding, as a means to an end.

This is why the effect is called overjustification. The person has an intrinsic reason to do the thing. The product supplies an additional, external reason. The external reason is now sufficient to explain the behavior, so the internal one is discounted. And when the external reason is withdrawn, the behavior falls below where it started, because the internal reason has been discounted away.

#### REPLICATION STATUS: THIS ONE HOLDS, AND IT IS QUANTIFIED

Unlike several effects in this chapter, overjustification has been subjected to a large meta-analysis with an explicitly adversarial history. Deci, Koestner and Ryan (1999), *A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation*, *Psychological Bulletin* 125(6), 627-668, synthesised **128 studies**. From the abstract, verbatim:

"*As predicted, engagement-contingent, completion-contingent, and performance-contingent rewards significantly undermined free-choice intrinsic motivation (d = –0.40, –0.36, and –0.28, respectively), as did all rewards, all tangible rewards, and all expected rewards.*"

And, crucially for design:

"*Positive feedback enhanced both free-choice behavior (d = 0.33) and self-reported interest (d = 0.31). Tangible rewards tended to be more detrimental for children than college students, and verbal rewards tended to be less enhancing for children than college students.*"

Two honesty notes. The meta-analysis is by the authors of self-determination theory, and it was written explicitly to rebut a prior meta-analysis (Cameron and Pierce, 1994) that had found otherwise. A reader should weigh that. The countervailing consideration is that the effect sizes are consistent in sign across

several reward contingencies and that the *pattern* of moderators is theoretically coherent rather than arbitrary: **tangible, expected, contingent rewards undermine; verbal positive feedback enhances.** Whether the 1999 meta-analysis survives a modern publication-bias-corrected re-analysis is not resolved here: CITATION NEEDED.

### THE ARGUMENT AGAINST STREAKS THAT NOBODY MAKES

The popular critique of streaks and achievements is that they are *addictive*. That critique is weak: it leans on the operant analogy that section 3.5 has just conceded is an analogy, and a defender can answer it by observing that the user wanted the habit.

The overjustification argument is harder to answer, and it runs in the opposite direction.

Take the paradigm case that Chapter 1 and the frameworks both use: a language-learning application, a user who genuinely wants to study daily, and a streak (**C-01**). The valence analysis in PROJECT.md declares this the *ethical* implementation. The user's baseline motive is intrinsic; they want to speak the language. The streak supplies an extrinsic, expected, engagement-contingent reward, which is precisely the class Deci et al. report as most strongly undermining ($d = -0.40$). The predicted long-run consequence is not that the user stops using the product. It is that their *reason* for using it migrates from the language to the number, and that when the number breaks, the behavior collapses further than it would have if the number had never existed. **The streak may not have scaffolded the intrinsic motive. It may have replaced it.**

Three taxonomy consequences follow.

**C-01 and C-09 carry a hidden cost even in their ethical implementation.** The Behavioral Cost Index scores the language-app streak at Emotion 1. Overjustification suggests that is too generous, because the cost is not felt at the time of the interaction; it is a deferred liability against the user's own motivation. That is the shape of **Behavioral Debt** (frameworks/03-genome-dna-debt-gravity.md), and it is a novel instance of it: a debt incurred not against the user's trust but against their intrinsic interest. Proposed here.

**C-04 (Streak Insurance) is worse than the frameworks already say.** The pattern manufactures an anxiety and sells the cure (section 3.10). Overjustification adds that the anxiety is attached to a token that has, over time, been *displacing* the user's real reason for being there.

**The ethical alternative is specifiable, and Deci et al. name it.** Positive verbal feedback *enhanced* free-choice behavior ($d = 0.33$) where tangible contingent reward undermined it. A product that wants to support a habit without corroding the motive behind it should tell the user what they achieved *in the domain they care about* rather than issue a token whose value the product controls. "You can now read a menu in Spanish" is informational feedback. "You have a 43-day streak" is a contingent extrinsic reward. The taxonomy has no pattern for the former, and it should. This is a gap in Family I, recorded in the Open Questions.

### MECHANISM TO PATTERN IDS

| MOTIVATIONAL MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|
| Expected, contingent extrinsic reward discounts the intrinsic motive (Lepper et al. 1973; Deci et al. 1999) | Streaks, achievements, tiers, badges | C-01, C-09, C-10, D-10 |
| Reward magnitude tied to consecutive engagement | Escalating rewards forfeited on lapse | C-12, C-03 |
| The manufactured motive is then monetised | Paid repair of the token the product created | C-04 |
| Status conferred and scarcity set by the platform | Platform-controlled distinction | D-10, D-06 |
| Informational feedback *enhances* intrinsic motivation (d = 0.33) | Not currently a taxonomy pattern | **gap in Family I** |

**The System 1 tie.** Overjustification is the one mechanism here that does not suppress deliberation in the moment. It operates on the *content* of the user's deliberation over months, by changing the reason they would give if asked. A chapter that described only momentary manipulation would miss the slowest and least reversible harm in the taxonomy.

---

## 3.7 Loss aversion and the endowment effect

#### THE STURDIEST THING IN THE CHAPTER

Kahneman and Tversky (1979), *Prospect Theory: An Analysis of Decision under Risk*, *Econometrica* 47(2), 263-291, defines value on gains and losses relative to a reference point rather than on final states, and states that "the value function is normally concave for gains, commonly convex for losses, and is generally steeper for losses than for gains." That final clause is the formal statement of loss aversion.

It replicates. Ruggeri et al. (2020), across **4,098 participants in 19 countries and 13 languages**, report that the patterns of prospect theory "replicated for 94% of items, with some attenuation," and that twelve of thirteen theoretical contrasts replicated. In a document required by its own constitution to report replication status, this is the one place where the underlying science can be leaned on hard, and section 3.11 treats it as the reference point against which everything else in the chapter is scored.

Chapter 1 has already used loss aversion in Chain 4 (price anchoring) and Chain 5 (loss-framed streak messaging), and this chapter does not restate that. What follows is the part Chapter 1 did not do: the **endowment effect**, and the systematic mapping.

#### THE ENDOWMENT EFFECT

The endowment effect is the finding that people demand more to give up an object than they would pay to acquire it. Kahneman, Knetsch and Thaler (1990), *Experimental Tests of the Endowment Effect and the Coase Theorem*, *Journal of Political Economy* 98(6), 1325-1348, report it directly. From the abstract, verbatim:

*"Contrary to theoretical expectations, measures of willingness to accept greatly exceed measures of willingness to pay. This paper reports several experiments that demonstrate that this 'endowment effect' persists even in market settings with opportunities to learn. Consumption objects (e.g., coffee mugs) are randomly given to half the subjects in an experiment. Markets for the mugs are then conducted. The Coase theorem predicts that about half the mugs will trade, but observed volume is always significantly less. When markets for 'induced-value' tokens are conducted, the predicted volume is observed, suggesting that transactions costs cannot explain the undertrading for consumption goods."*

The induced-value control is what makes the result hard to dismiss. When the traded object is a token with a known cash value, volume matches theory. When it is an object of consumption, it does not. The asymmetry is not a friction in the market; it is a fact about how a good is valued once it is *yours*.

The relationship between the two effects should be stated so that it is not overstated: the endowment effect is what the loss-aversion asymmetry predicts when a person is asked to give up something they hold, because giving it up is coded as a loss against the reference point their possession established.

#### THE MAPPING

Once loss aversion and endowment are on the table, an entire class of retention machinery becomes legible as one mechanism, and the mechanism has a name: **manufacture an endowment, then threaten it.**

**C-01 (Streak).** The streak's value derives entirely from its fragility, which is the taxonomy's own definition of the pattern. Nothing is actually lost by missing a day, and nothing tangible was ever gained by not missing one. The streak is an endowment the product created out of nothing, gave to the user, and then made destructible. Once the user holds it, prospect theory says they will weight its loss more heavily than they ever weighted its acquisition, so "you will lose your streak" draws on a value asymmetry the product itself installed. It is the cleanest example in the taxonomy of a pattern whose entire power is borrowed from a verified effect.

**C-12 (Reward Escalation).** The same mechanism with a ramp: reward magnitude rises with consecutive engagement, so the endowment grows, so the prospective loss from lapsing grows. The product has increased the *cost of leaving* without increasing the *value of staying* by a single unit, which in the Behavioral Influence Model's terms is the definition of the manipulative move.

**C-10 (Tier Progression)** and **D-12 (Reputation Lock-in)** are endowments in status and standing, non-portable by construction, so their full value is forfeited on exit.

**D-09 (Missing-Out Signal).** Loss aversion applied to counterfactuals. It surfaces what the user *did not see* and frames absence as loss. The user has lost nothing; they simply were not present. The pattern installs a reference point in which presence was the baseline, so that absence is coded as a loss rather than as a neutral non-event. Emotionally this is FOMO (section 3.10); structurally it belongs here.

**H-07 (Downgrade Penalty).** The most economically explicit of the family. A user reducing spend is confronted not with the loss of the feature they are giving up but with the loss of unrelated function they had come to rely on. The downgrade is engineered to be evaluated as a *loss* from the current holding

rather than as a *gain* relative to no subscription at all, and the value function says those framings are not equivalent even when the final states are identical.

**H-06 (Data Hostage)** and **H-02 (Exit Offer)** complete the set at departure. The first threatens an endowment the user actually created. The second has a genuinely contested valence, and the contest turns on one question: does the discount arrive *instead of* the cancel button, or *beside* it? Beside it, this is a better deal offered at a moment of high information. Instead of it, it is friction wearing the costume of generosity, and the pattern is **H-01** in disguise.

### THE EXCEPTION THAT PROVES THE MECHANISM

**I-08** (Reversible Commitment) and **I-03** (Frictionless Exit) turn the same machinery around. Both work by *lowering* the perceived cost of a prospective loss: they tell the user, credibly, that what they are about to do can be undone. The Behavioral Cost Index scores **I-03** at zero on every axis and notes that it still serves a business goal, "a user who trusts that they can leave is more willing to arrive."

That sentence is a loss-aversion claim, and it is the ethical use of the same effect. A prospective user evaluating a subscription is evaluating a prospective loss of money, of freedom, and of the hassle of extraction. Family I is not asking companies to ignore prospect theory. It is asking them to apply it at the entrance rather than at the exit.

### MECHANISM TO PATTERN IDS

| LOSS-AVERSION MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|
| Value is defined against a reference point (K&T 1979) | Reference price shown beside the offer | **F-05**, **F-06** |
| Losses weigh more than equivalent gains (K&T 1979; replicated Ruggeri et al. 2020) | Lapse framed as loss, not as a non-event | **C-01**, **C-14**, **D-09** |
| Endowment raises willingness to accept above willingness to pay (KKT 1990) | Manufacture a holding, then make it destructible | **C-01**, **C-12**, **C-10**, **D-12** |
| Non-portable endowment forfeits fully on exit | Status, reputation, and content locked to the platform | **D-07**, **D-12**, **H-06**, **G-09** |
| Downgrade evaluated as loss from current state | Punitive removal of unrelated function | **H-07**, **H-10** |
| Prospective loss *reduced* honestly | Reversibility, portability, symmetric exit | **I-03**, **I-07**, **I-08**, **I-11** |

**The System 1 tie.** Loss aversion marks the boundary of the organizing claim, because it is not a failure of System 2: a user who knows about it still feels it. Patterns that merely exploit the asymmetry are therefore often *persuasive* rather than manipulative, since disclosure does not defeat them. What tips them into manipulation is the **manufacture of the endowment**. A user told plainly that the streak is a number the company invented, that it protects nothing, and that its loss costs them nothing real, is a user whose System 2 has something to work with. That is the disclosure **C-04** cannot survive.

## 3.8 Goals, progress, and resumption

#### ENDOWED PROGRESS

Nunes and Drèze (2006), *The Endowed Progress Effect: How Artificial Advancement Increases Effort*, *Journal of Consumer Research* 32(4), 504-512, document that people given artificial advancement toward a goal persist more toward reaching it. From their abstract, verbatim:

"*By converting a task requiring eight steps into a task requiring 10 steps but with two steps already complete, the task is reframed as one that has been undertaken and incomplete rather than not yet begun. This increases the likelihood of task completion and decreases completion time.*"

Two things must be said about this citation and both are enforced by PROJECT.md.

**The mechanism is framing, not sunk cost.** The abstract states that the effect "appears to depend on perceptions of task completion rather than a desire to avoid wasting the endowed progress." Product writing routinely explains endowed progress as sunk-cost avoidance. The authors explicitly say it is not.

**The widely circulated completion percentages are not printed here.** They are not confirmed against the primary text; the abstract does not contain them and does not mention a car wash at all. This document cites the 8-versus-10 manipulation and the direction of the effect, which is what the retrieval supports, and nothing else.

**B-03** (Endowed Progress) is the taxonomy's pattern, and it is one of the few whose lineage is fully verified in TAXONOMY.md. Its interface expression is the onboarding checklist that opens at 20 percent complete because the user created an account, or the loyalty card that arrives with two stamps already on it.

#### GOAL GRADIENT

The goal-gradient hypothesis holds that effort intensifies as a goal draws nearer. Kivetz, Urminsky and Zheng (2006), *The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention*, *Journal of Marketing Research* 43(1), 39-58, tested it in humans, in the field, and their abstract is worth quoting at length because it does four separate jobs:

"*The key findings indicate that (1) participants in a real café reward program purchase coffee more frequently the closer they are to earning a free coffee; (2) Internet users who rate songs in return for reward certificates visit the rating Web site more often, rate more songs per visit, and persist longer in the rating effort as they approach the reward goal; (3) the illusion of progress toward the goal induces purchase acceleration (e.g., customers who receive a 12-stamp coffee card with 2 preexisting 'bonus' stamps complete the 10 required purchases faster than customers who receive a 'regular' 10-stamp card); and (4) a stronger tendency to accelerate toward the goal predicts greater retention and faster reengagement in the program.*"

Note what finding (3) is. It is an independent replication of the endowed-progress manipulation, in a different laboratory, with a different reward, in the field, and it is the *same* structure: a card requiring

more nominal steps but pre-credited, against a card requiring fewer nominal steps and starting empty. The work demanded is identical. The behavior is not.

Note also finding (4), which is the one the product industry actually monetises: acceleration toward a goal *predicts retention*. The goal gradient is not merely a curiosity about effort; it is a retention instrument, and it was published as one, in a marketing journal, in 2006.

Kivetz et al. themselves attribute the original hypothesis to the behaviorist Clark Hull, whose rats "ran progressively faster as they proceeded from the starting box to the food." The animal-to-human extrapolation problem that dogs the operant literature (section 3.5) **does not apply here**, because Kivetz et al. did not extrapolate. They ran the human experiments.

**C-05** (Goal Gradient) is the taxonomy's pattern. **B-02** (Completion Meter) is its onboarding-stage relative: a progress bar converts an open-ended task into a closable one, which is the precondition for a gradient to exist at all. There is no gradient without a visible end.

#### RESUMPTION, AND THE CORRECTION THAT HAS TO PROPAGATE

Now the correction, which Chapter 1 introduced in Chain 6 and which this section is the proper home for.

Every UX text that discusses progress bars, incomplete profiles, and unfinished onboarding attributes the pull of the incomplete to the **Zeigarnik effect**: the claim that interrupted tasks are better *remembered* than completed ones.

The Zeigarnik effect does not replicate. Ghibellini and Meier (2025), *Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects*, *Humanities and Social Sciences Communications* 12, article 962, report a pooled ratio of recall for interrupted versus completed tasks of **0.99** (k = 38), a proportion of interrupted tasks among all recalled of 49.16 percent (k = 13), which is chance, and a small d(z) of 0.15 (k = 8). Their conclusion, quoted: "The current findings do not support a memory advantage for interrupted tasks when situational influences and individual differences are not accounted for." And their summary line: "The Ovsiankina effect represents a general tendency, whereas the Zeigarnik effect lacks universal validity."

What *does* replicate is the **Ovsiankina effect**: the tendency to **resume** an interrupted task. Pooled resumption rate **67.0 percent** (k = 21).

The taxonomy is not weakened by this. It is strengthened, and the reason is worth stating carefully because it is the clearest instance in this document of replication honesty producing a *better* theory rather than a smaller one.

A progress bar does not work by making the user remember the unfinished profile. Nobody has ever claimed to be haunted by the memory of a 60-percent-complete LinkedIn profile. What the unfinished state does is exert a pull toward **resumption** when the user is next in a position to act. The behavior that the pattern actually produces, and that companies actually measure, is *return and completion*. That is Ovsiankina's phenomenon, not Zeigarnik's. The product literature has spent two decades citing the wrong effect for a behavior that a neighbouring, better-evidenced effect explains exactly.

**C-13** (Completion Tension) is therefore grounded in a *stronger* effect than the one product writing habitually cites, and TAXONOMY.md already records the correction in its lineage table.

### THE THREE PROGRESS PATTERNS ARE DIFFERENT MECHANISMS AND SHOULD NOT BE CONFLATED

This is the section's original contribution to the taxonomy, and it matters because the three are constantly treated as one.

| PATTERN | MECHANISM | VERIFIED SOURCE | WHAT IT PREDICTS |
|---|---|---|---|
| **B-03** Endowed Progress | Reframing a not-yet-begun task as an already-undertaken one | Nunes & Drèze 2006; independently reproduced as finding (3) in Kivetz et al. 2006 | Higher completion, faster completion |
| **C-05** Goal Gradient | Effort intensifies with proximity to a visible goal | Kivetz et al. 2006 (human, field) | Acceleration near the end; predicts retention |
| **C-13** Completion Tension | Disposition to resume an interrupted task | Ovsiankina, per Ghibellini & Meier 2025 (pooled resumption 67.0%) | Return and finish, not superior recall |

The distinction is operational, not academic. If a designer wants people to *start*, endowed progress is the instrument, and the evidence says it works by changing the perceived nature of the task. If a designer wants people to *finish faster*, the goal gradient is the instrument, and the evidence says it works by proximity. If a designer wants people to *come back*, resumption is the instrument, and the evidence says nothing about memory at all. Citing "Zeigarnik" for all three, which is the industry norm, conceals the fact that these are three different levers with three different failure modes.

### MECHANISM TO PATTERN IDS

| GOAL MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|
| Artificial advancement reframes the task (Nunes & Drèze 2006) | Checklist that opens partly complete; pre-stamped card | **B-03**, **B-02** |
| Effort intensifies near a visible goal (Kivetz et al. 2006) | Progress made visible precisely as the end nears | **C-05**, **C-10**, **C-12** |
| Acceleration predicts retention (Kivetz et al. 2006, finding 4) | Reward programs designed around the gradient | **C-05**, **C-03**, **E-01** |
| Disposition to resume an interrupted task (Ovsiankina; Ghibellini & Meier 2025) | Deliberately visible unfinished state | **C-13**, **B-08**, **A-03**, **A-07** |
| A goal is *needed* for a gradient to exist | Open-ended task converted into a closable one | **B-02**, **B-01** |
| Progress made honest | Completion that reflects real value, not product-defined tasks | **gap in Family I** |

**The System 1 tie.** All three operate by changing the user's *representation* of a task rather than the task itself, and all three work best when the user is not asking the System 2 question, which in every case is the same one: *has the amount of work I must do changed?* It has not. In **B-03** it provably has not, because the manipulation holds the work constant by construction.

## 3.9 Social cognition

### THREE MECHANISMS, AND THE HONEST STATE OF THE EVIDENCE

Family D of the taxonomy recruits other people, and the user's self-image, as the engine of engagement. The psychological substrate is social cognition, and the practitioner canon for it is Cialdini's *Influence*.

Two facts about that canon must be stated before it is used.

**Cialdini now lists seven principles, not six.** As enumerated on his own site: reciprocity, scarcity, authority, consistency, liking, social proof, and unity. **Unity** is the addition, described as shared identity, the sense that the persuader is "one of us." A different page on the same official site still says "six major factors," carrying a footnote that reads "This was created before the addition of the 7th Principle of Unity." Cialdini's own site is therefore internally inconsistent on the count, which is a small but telling illustration of how a practitioner canon propagates. This document cites the seven-principle page.

**Replication is mixed and principle-dependent.** The verified dossier records the position: reciprocity and social proof are well supported; some of the specific studies in the book come from the era and style of social psychology that has replicated poorly. "Cialdini's principles" must therefore not be presented as uniformly settled science, and this chapter does not present them that way.

The chapter can do better than a general caveat, because a specific and instructive case was retrieved.

### THE SOCIAL-PROOF CASE, WORKED ALL THE WAY THROUGH

**The original.** Goldstein, Cialdini and Griskevicius (2008), *A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels*, *Journal of Consumer Research* 35(3), 472-482. Two field experiments on hotel signage. From the abstract, verbatim: "Appeals employing descriptive norms (e.g., 'the majority of guests reuse their towels') proved superior to a traditional appeal widely used by hotels that focused solely on environmental protection." Appeals were most effective when the norm described behavior in the setting closest to the guest's own circumstances ("the majority of guests *in this room*"), which the authors call **provincial norms**.

This is the paper behind every "127 people are looking at this hotel right now" message on the internet. **D-01** (Social Proof Counter) is the taxonomy's pattern for it.

**The replication.** Bohner and Schlüter (2014), *A Room with a Viewpoint Revisited: Descriptive Norms and Hotel Guests' Towel Reuse Behavior*, *PLOS ONE* 9, e104086, ran the same design in two German hotels (Study 1, N = 724; Study 2, N = 204), comparing a descriptive-norm message stating that 75 percent of guests had reused their towels against a standard environmental appeal. In Study 1 the descriptive-norm message did not outperform the standard appeal. In Study 2, the standard environmental message *outperformed* the descriptive-norm messages. Baseline reuse rates in the German hotels were far higher than in the original US research.

**What to conclude, honestly.** Not that social proof is fake. The mechanism is real, is old, and is not seriously in doubt as a psychological phenomenon. What the replication shows is that the *effect of a descriptive norm message on behavior* is highly contingent on the base rate of the behavior in that population. Where a behavior is already near ceiling, telling people that most others do it adds nothing,

and may even underperform a straightforward appeal, plausibly because it reframes a moral act as a conformity act. That is a boundary condition, and boundary conditions are what distinguish a science from a canon.

For the taxonomy this is a genuinely useful result, because it makes a *prediction about valence*. **D-01** is at its most persuasive, and its least manipulative, when the number displayed is true and the base rate it reports is real. It is at its most manipulative when the number is invented, which is **F-15** (Scarcity Claim) and Brignull's "fake social proof," or when the base rate is unrepresentative and the display is engineered to suggest otherwise. Mathur et al. (2019) found social proof and scarcity to be measurable at scale precisely because the fabricated versions leave detectable traces, and their crawl of roughly 11,000 shopping sites reports "Activity Messages" and "Testimonials of Uncertain Origin" as named types.

#### RECIPROCITY, CONFORMITY, AND THE OBLIGATION PATTERNS

**Reciprocity** is the principle that a received favour creates a felt debt. It is one of the two principles the dossier records as well supported. Its interface expression is **D-02** (Reciprocity Prompt), and it is the psychological engine of the follow-back, the profile view notification, and the "X endorsed you" message. The debt is not to the platform; it is to another user, which is what makes the mechanism so efficient. The platform does not need to motivate the return action. It only needs to make the imbalance visible.

**D-03** (Read Receipt) is the same logic in its most refined form and deserves particular attention, because it is the purest case in the taxonomy of a feature that manufactures an obligation out of information. A read receipt converts a message into a debt by making one party's attention visible to the other. The recipient has now been *seen* to have seen. Not replying has become an act rather than an omission. Nothing has been added to the conversation; a fact about it has been published, and the publication is what does the work. Its ceiling of 4 is earned, and its ethical treatment is a matter of who controls the disclosure.

**Conformity** is the disposition to align one's behavior with a group's. Its expressions are **D-06** (Comparative Ranking), which converts use into competition, and **D-11** (Group Obligation), which arranges matters so that a lapse harms other people. **D-11** deserves to be called out as the most severe pattern in the family's logic, because it takes the guilt mechanism of **C-14** and outsources its enforcement to the user's friends. The company no longer has to send the guilt message. The company has arranged for a peer to send it.

**Unity**, Cialdini's seventh, is the taxonomy's **D-07** (Identity Investment) and **D-08** (Parasocial Hook), and it is where social cognition meets the Identity node of the Behavior Loop. A user who has come to understand themselves as a member of a product's community is a user for whom leaving is a change of self rather than a change of software. Chapter 1 argued that Identity is the node at which behavioral design becomes both most valuable and most dangerous. Family D is where that node is engineered.

#### MECHANISM TO PATTERN IDS

| SOCIAL MECHANISM | EVIDENTIAL STATUS | INTERFACE EXPLOITATION | PATTERN IDS |
|---|---|---|---|
| Descriptive norms shift behavior (Goldstein et al. 2008) | Replicates *conditionally*; fails where base rate is at ceiling (Bohner & Schlüter 2014) | Activity counters, "most guests," "127 viewing" | **D-01**, **F-15** |
| Fabricated norms | Not a psychological finding; a deception | Invented counts and testimonials | **F-15**, **F-14**, **D-01** (manipulative) |
| Reciprocity creates a felt debt (Cialdini; dossier records as well supported) | Supported | Follow-backs, endorsements, visible imbalance | **D-02**, **E-01** |
| Visibility of attention creates obligation | **CITATION NEEDED** for the specific claim | Read receipts, typing indicators, presence | **D-03**, **D-04**, **A-11** |
| Conformity and comparison | Mixed; principle-dependent | Leaderboards, ranks, public counts | **D-05**, **D-06** |
| Obligation outsourced to peers | Follows from reciprocity + guilt | Shared streaks and group goals | **D-11**, **C-14** |
| Unity and shared identity (Cialdini's 7th) | Newest principle; least tested | Identity investment, parasocial attachment | **D-07**, **D-08**, **D-12** |

**The System 1 tie.** Social inference is fast, automatic, and does not report its workings, which is the textbook description of Type 1 processing. A user who sees that 127 people are viewing a hotel room does not audit the counter. They feel the room is in demand. The claim is not that they *could* not check; it is that checking is a System 2 operation, and the counter is placed at a moment engineered to be a System 1 one.

---

## 3.10 Emotion and affect

#### AFFECT IS A LEVER, AND IT IS THE CHEAPEST ONE

The Behavioral Cost Index's seventh axis is **Emotion**, and its diagnostic question is the bluntest in the instrument: *did the product make the user feel bad on purpose?*

The question is worth having as an axis because affect is not a by-product of behavioral design. It is frequently the mechanism. A pattern that cannot make the user *do* anything can still make them *feel* something, and the feeling will do the work.

#### ANXIETY, MANUFACTURED AND THEN SOLD

The paradigm case is the one the frameworks already use, and this section is where its psychology is named.

**C-01** (Streak) creates a token whose value derives from its fragility. **C-14** (Guilt Messaging) frames a lapse as a personal or relational failure. **C-04** (Streak Insurance) sells a repair for the broken streak.

The Behavioral Cost Index scores that combination at **Emotion 5**, and the scoring rule for a 5 is precise: "cost imposed, user is unaware, awareness would change their behavior, *and the design works precisely because of that unawareness*." The justification the framework gives is the one this chapter endorses: the product manufactures an anxiety that has no referent in reality, since nothing is genuinely lost by

missing a day, and then sells the remedy for it. It works *because* the user has not noticed that the thing they are afraid of losing was created by the company that is now charging them to protect it.

Two mechanisms from earlier in this chapter converge on that sentence, and neither is redundant. Loss aversion (3.7) explains why the manufactured endowment can generate an emotion at all. Overjustification (3.6) explains why, over time, the manufactured token may be *displacing* the intrinsic reason the user was there. The anxiety is not merely unpleasant; it is anxiety about the loss of a thing that has quietly become the user's reason for acting.

### FOMO, AND WHAT CAN HONESTLY BE SAID ABOUT IT

The fear of missing out is the affective face of **D-09** (Missing-Out Signal), which section 3.7 analysed structurally as a reference-point manipulation.

Przybylski, Murayama, DeHaan and Gladwell (2013), *Motivational, emotional, and behavioral correlates of fear of missing out*, *Computers in Human Behavior* 29(4), 1841-1848, developed and evaluated a self-report measure of FOMO and reported that it was associated with lower need satisfaction, lower mood, and lower life satisfaction, and was linked to higher levels of social media engagement. The bibliographic record for this paper was verified against Crossref. **The specific findings summarised in the preceding sentence were taken from search-result summaries and secondary descriptions, not from a retrieved primary text**, and are therefore marked: CITATION NEEDED for the specific correlational claims, pending retrieval of the article itself. The citation is real; the results as stated here are not yet primary-verified. This is exactly the distinction the Citation Law exists to enforce, and it is left visible rather than smoothed over.

What can be said without any of that, on structural grounds alone, is the part that matters for the taxonomy: **D-09** does not report information the user lacked. It converts a non-event into a loss by supplying a reference point in which the user's presence was the baseline. Whether a user experiences that as distress is an empirical question. Whether the design constructed the loss is not.

### ANTICIPATION, AND THE LIMITS OF WHAT CAN BE CLAIMED

Not every affective pattern manufactures a bad feeling. **A-03** (Badge Signal) and **A-07** (Unread Debt) exploit a mild, ambient discomfort with unresolved items, and section 3.8 gives it its correct name: the disposition to resume. **A-05** (Pull-to-Refresh) exploits anticipation of an uncertain payoff, and section 3.5 is explicit that the operant reading of that is an analogy. The honest position on the affective side of Family A is therefore weaker than on Families C and D. The claim that badge counts and pull-to-refresh generate a specific, measurable affective state in users has no retrieved primary source in this document: CITATION NEEDED.

### EMOTION AS THE AXIS THAT SEPARATES PERSUASION FROM MANIPULATION

There is a general result available here, and it is the section's contribution.

Recall the Disclosure Test from the Pattern DNA framework: narrate the pattern's full causal chain to the user in plain language, and see whether it still works. Apply it, and a striking asymmetry appears.

Patterns that work through **perception, memory load, and default inertia** are often *not* defeated by disclosure. A user who knows what an anchor is still anchors. A user who knows about defaults still keeps them, because the default's power comes from the cost of acting, not from a false belief.

Patterns that work through **manufactured emotion** are almost always defeated by disclosure. Tell the user that the anxiety they feel about their streak was designed, that the streak protects nothing, and that its loss costs them nothing real, and the emotion has nowhere to attach. That is why the Pattern DNA table names *Emotion* as the link that breaks on disclosure for **C-04**.

The generalisation, proposed here: **manufactured affect is the most fragile of all the mechanisms in this chapter, and therefore the most dependent on concealment, and therefore the most reliably manipulative.** A pattern whose power resides in an emotion the product created is a pattern that cannot survive being explained. By the valence definition in PROJECT.md section 4, that is not a strong form of persuasion. It is the definition of manipulation.

MECHANISM TO PATTERN IDS

| AFFECTIVE MECHANISM | INTERFACE EXPLOITATION | PATTERN IDS | BCI AXIS |
|---|---|---|---|
| Anxiety attached to a manufactured endowment | Streak whose value is its fragility | **C-01**, **C-12** | Emotion |
| Guilt framing of a lapse | Copy that codes non-use as failure | **C-14**, **H-03**, **F-08** | Emotion, Trust |
| The manufactured anxiety is monetised | Purchasable repair | **C-04**, **H-10** | Emotion, Money |
| Absence framed as loss | Surfacing what was missed | **D-09**, **A-10** | Emotion |
| Shame at the point of declining | Decline option worded to shame | **F-08**, **H-03** | Emotion, Trust |
| Ambient tension of the unresolved | Badges and unread counts | **A-03**, **A-07**, **C-13** | Attention, Emotion |
| Anticipation of an uncertain payoff | Manual gesture, uncertain return | **A-05**, **F-16** | Emotion |
| Affect *not* weaponised | Sufficiency, honest exit, unused-value alerts | **I-05**, **I-03**, **I-10** | Emotion |

**The System 1 tie.** Affect is the fastest of all the systems in this chapter and the least available to introspection. An emotion arrives before the reasoning that would evaluate it, which is precisely why a manufactured emotion is such an efficient way of ensuring that the reasoning never runs.

---

## 3.11 The limits of this substrate

A chapter on behavioral science written in 2026 that does not address the replication crisis is not credible, and would in any case be arguing against itself: this document's central methodological commitment is that replication status is part of the citation. This section is where that commitment is discharged in full, and it is written as a feature of the argument rather than as a disclaimer at the end of it.

THE SCALE OF THE PROBLEM

The Open Science Collaboration (2015), *Estimating the reproducibility of psychological science*, *Science* 349(6251), aac4716, replicated 100 experimental and correlational studies from three psychology journals using high-powered designs and original materials where available. From the abstract, verbatim:

"*Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results.*"

Ninety-seven percent to thirty-six percent. That is the number that should be held in mind while reading any chapter of this kind, including this one.

The single most instructive individual case is ego depletion, because of how central it had become. Hagger et al. (2016), *A Multilab Preregistered Replication of the Ego-Depletion Effect*, *Perspectives on Psychological Science* 11(4), 546-573, coordinated **23 laboratories** and **2,141 participants** in a preregistered replication of a standardised sequential-task ego-depletion protocol. From the abstract, verbatim:

"*Meta-analysis of the studies revealed that the size of the ego-depletion effect was small with 95% confidence intervals (CIs) that encompassed zero (d = 0.04, 95% CI [–0.07, 0.15]).*"

An effect size of 0.04 with a confidence interval straddling zero, from a preregistered multi-lab design, is as close to a null result as this literature produces. And ego depletion was not a marginal finding. It was the theoretical basis for the entire "willpower is a depletable resource" edifice, and by extension for a large body of design writing which argues that users make worse decisions late in a session because their self-control is exhausted.

**This document does not make that argument, and this is the place to say so explicitly.** The claim that "decision fatigue" explains why users accept unfavourable terms at the end of a long flow is not available to us. It is exactly the sort of claim that the persuasive-design literature repeats without checking, and it rests on a body of work whose flagship effect did not survive a preregistered multi-lab test.

What we say instead, and what the evidence supports, is narrower and stronger: **users accept unfavourable terms at the end of a long flow because attention is depleted by task load in a way Simons and Chabris demonstrated, because working memory has a capacity of about four chunks (Cowan, 2001), and because a switching cost is paid for every interruption (Rubinstein et al., 2001).** Those are load claims, not willpower claims, and they do not depend on the resource model of self-control. Replacing an ego-depletion story with a cognitive-load story costs this document nothing and gains it a defensible foundation.

#### TIERING THE EVIDENCE IN THIS CHAPTER

Every mechanism in this chapter is placed in one of four tiers. This table is the chapter's most important single artifact, and the honest reader is invited to check it.

| TIER | MEANING | MECHANISMS | PATTERNS THAT REST ON THEM |
|---|---|---|---|
| **Load-bearing and safe** | Verified primary source, plus independent replication at scale | Prospect theory / loss aversion (K&T 1979; Ruggeri et al. 2020, 4,098 participants, 19 countries, 94% of items). Endowment effect (KKT 1990, with induced-value control). Working-memory capacity (Cowan 2001, revising Miller 1956). Inattentional blindness (Simons & Chabris 1999, 46% missed). Task-switch cost (Rubinstein et al. 2001). Resumption / Ovsiankina (Ghibellini & Meier 2025, pooled 67.0%). Endowed progress (Nunes & Drèze 2006, independently reproduced by Kivetz et al. 2006). Goal gradient (Kivetz et al. 2006, human field experiments). Overjustification (Lepper et al. 1973; Deci et al. 1999, 128 studies). Defaults (Johnson & Goldstein 2003). | C-01, C-12, D-09, H-07, F-04, F-17, F-18, G-02, G-12, B-02, B-03, C-05, C-13, C-09, A-04, A-06 |
| **Contested, use with the caveat printed** | Real phenomenon, disputed magnitude or boundary conditions | Social proof / descriptive norms (Goldstein et al. 2008, *conditionally* replicated; reversed in Bohner & Schlüter 2014 where base rate was at ceiling). Cialdini's principles generally (seven now, not six; replication mixed and principle-dependent; his own site is self-inconsistent). Nudge as a body (pooled effect disputed under publication-bias correction; **no nudge effect size appears anywhere in this document**). Dual-process theory itself (defended by Evans & Stanovich 2013, but as *types* of processing, not as two literal systems). | D-01, D-02, D-05, D-06, F-15, and the framing of section 3.1 |
| **Analogy, not finding** | The mechanism is established in a different domain and has been transported | Variable-ratio reinforcement in human app use (Ferster & Skinner 1957 is animal-laboratory work; the transfer is an extrapolation across species, setting, and response class). The Hook Model (Eyal 2014; a practitioner framework with no primary empirical base of its own). | C-02, A-05, C-07, C-03 |
| **Dead** | The named effect does not survive meta-analysis | The **Zeigarnik memory effect**: pooled recall ratio **0.99** (k = 38); interrupted tasks are **not** better remembered | None. C-13 was re-grounded on Ovsiankina; no pattern in this document rests on ego depletion. |

| TIER | MEANING | MECHANISMS | PATTERNS THAT REST ON THEM |
|---|---|---|---|
| | | (Ghibellini & Meier 2025). Ego depletion as a resource model: **d = 0.04, 95% CI [–0.07, 0.15]** across 23 labs and 2,141 participants (Hagger et al. 2016). | |

### WHAT FOLLOWS FOR A TAXONOMY BUILT ON THIS SUBSTRATE

Four consequences.

**The tiering redistributes the taxonomy's confidence, and not where the industry assumes.** The families the popular critique attacks hardest, Family C and the variable-reward story, rest on the *weakest* tier. The families it barely mentions, Families F and G and the whole machinery of defaults, memory load, and reference points, rest on the *strongest*. A regulator who reads the popular literature will regulate the slot machine and miss the checkout flow.

**A dead effect does not kill a pattern.** The Zeigarnik result was a genuine loss of a citation and no loss at all of a phenomenon. Progress bars still work; incomplete profiles still get completed. What changed is *which mechanism we are entitled to name*, and the correct one turned out to be better evidenced than the wrong one. A taxonomy organised by mechanism can absorb this. A taxonomy organised by folklore cannot, because it has no way to tell which of its entries just lost their basis.

**The frameworks inherit the uncertainty and must say so.** The Behavioral Influence Model places Reward Uncertainty in the numerator of persistence, and section 3.5 has shown that this term is the strongest form of the operant extrapolation. The framework may propose it; it may not present it as established. Its own limitations section already concedes the model "is not empirically validated," and that concession is now specific: **Reward Uncertainty is the BIM's weakest link, and it is the one a hostile reader should attack first.**

**The replication crisis is an argument *for* the Behavioral Cost Index, not against it.** The BCI requires no contested effect to be true. It scores the gap between the cost a user pays and the cost they understand, and that gap is an observable property of an interface, not a psychological hypothesis. A framework whose validity survives the collapse of half the literature it was built beside is a framework worth having.

### THE LIMITS OF THE LIMITS

This section does not claim that psychology is unreliable and that behavioral design therefore rests on nothing. That conclusion is popular, lazy, and refuted by the first tier of the table above: prospect theory replicated across 19 countries, and working-memory capacity is among the most robust quantitative findings in cognitive science.

Nor does it claim that this chapter's evidential base is complete. It is not. Twelve CITATION NEEDED markers are left visible in the body, one of them on a severity-5 pattern (**C-08**, Near Miss), and each is an instruction to a future editor: retrieve the primary source or cut the sentence. A visible gap beats an invisible fabrication.

### Key Takeaways

1. **Interfaces act on identifiable cognitive machinery, not on "the user."** An attentional bottleneck, a

salience-driven perceptual system, a working memory of roughly four chunks, a reinforcement-sensitive habit system, a motivational system corruptible by its own rewards, an asymmetric value function, a goal system that resumes what it left unfinished, a social-inference system, and an affective system that can be loaded on demand. Each maps to specific pattern IDs, and that mapping is this chapter's contribution.

1. **The organizing claim: every deceptive pattern is an attempt to keep the user in System 1 during a

decision that deserves System 2.** Deceptive design rarely argues with the deliberative system. It arranges for the deliberative system not to be consulted, by manipulating time available, working memory free, salience, and the cost in clicks of the deliberative path. *(Author's own formulation, proposed here.)*

1. **Working memory is the substrate of deliberation, and it is small.** Cowan (2001) revises Miller (1956)

downward to about four chunks where recoding is blocked, and Miller himself called seven "only a pernicious, Pythagorean coincidence." This makes the Behavioral Cost Index's Memory axis an instrument: **F-04** (Drip Pricing) does not make a demanding request of the user's arithmetic, it makes an impossible one.

1. **Attention is a filter that removes what it is not looking for.** 46 percent of observers missed a person

in a gorilla suit while counting basketball passes, and noticing fell further under load (Simons & Chabris, 1999). Disclosures do not have to be hidden to be unseen, which is what makes **G-06** (Setting Obscurity) defensible in court and effective in practice.

1. **Loss aversion is the sturdiest effect in the chapter and the engine of the retention families.** It

replicates across 4,098 participants in 19 countries (Ruggeri et al., 2020), and the endowment effect (Kahneman, Knetsch & Thaler, 1990) survives an induced-value control. The pattern common to **C-01**, **C-12**, **D-09**, and **H-07** is: **manufacture an endowment, then threaten it.** The endowment is manufactured; the aversion to losing it is real.

1. **The overjustification effect is the underused argument against gamification, and it is quantified.**

Expected, contingent, tangible rewards significantly undermine intrinsic motivation (Deci, Koestner & Ryan, 1999; 128 studies; d = –0.40 for engagement-contingent rewards), while positive verbal feedback *enhances* it (d = 0.33). So **C-01** and **C-09** may carry a deferred cost against the user's own motive even in their "ethical" implementation. That is a novel species of Behavioral Debt, proposed here.

1. **Zeigarnik is dead, Ovsiankina is alive, and the taxonomy is stronger for it.** Interrupted tasks are not

better remembered (pooled recall ratio 0.99). They are more likely to be *resumed* (pooled 67.0 percent). Progress bars and unfinished onboarding are resumption phenomena, and the right effect is better evidenced than the wrong one the industry has cited for two decades.

1. **Three progress patterns are three different mechanisms.** B-03 reframes a task not yet begun. C-05

intensifies effort near a visible end and predicts retention (Kivetz et al., 2006). C-13 exploits resumption. Conflating them under the single misapplied label "Zeigarnik" conceals three levers with three different failure modes.

1. **Social proof replicates conditionally, and the boundary condition is diagnostic.** The hotel-towel

descriptive norm (Goldstein, Cialdini & Griskevicius, 2008) failed to beat a standard environmental appeal in a German replication and was beaten by it in one study (Bohner & Schlüter, 2014), where the base rate was already high. D-01 is persuasive when the number is true and the base rate real, manipulative when either is fabricated.

1. **Manufactured affect is the most fragile mechanism and therefore the most reliably manipulative.**

Anchoring survives disclosure; defaults survive disclosure. An anxiety the product created does not. This is why the Pattern DNA table names *Emotion* as the link that breaks for C-04, and it generalises.

1. **The replication crisis redistributes confidence, and not where the critique expects.** Ninety-seven

percent of original studies were significant; thirty-six percent of replications were (Open Science Collaboration, 2015). Ego depletion did not survive 23 labs and 2,141 participants (d = 0.04, CI crossing zero; Hagger et al., 2016), so this document makes no "decision fatigue" argument and uses cognitive load instead. The families the popular critique attacks hardest rest on the weakest evidence; the families it ignores rest on the strongest.

---

## Open Questions

1. **Who actually coined "System 1" and "System 2," and does it matter?** The attribution to Stanovich and

West, and Kahneman's credit to them, is marked CITATION NEEDED in section 3.1 because the retrieved abstract of Stanovich and West (2000) does not contain the terms and the body was not retrieved. Retrieve the body of the target article and the relevant pages of *Thinking, Fast and Slow*, or drop the attribution entirely and cite Evans and Stanovich (2013) alone, which is sufficient for everything this chapter argues.

1. **What is the psychological basis of C-08 (Near Miss)?** It is a severity-5 pattern whose mechanism is

currently asserted rather than cited. The gambling literature on near-miss effects is the obvious place to look and was not retrieved in this pass. Until it is, the pattern entry is not defensible at its current ceiling.

1. **Does the overjustification effect survive a modern, publication-bias-corrected re-analysis?** Deci,

Koestner and Ryan (1999) is a 128-study meta-analysis written by the authors of the theory it supports, in explicit rebuttal of a competing meta-analysis. That is not disqualifying, but it is exactly the configuration that the replication era has taught us to check. If the effect attenuates substantially, the strongest argument in section 3.6 weakens, and the chapter should say so.

1. **Is the Reward Uncertainty term of the Behavioral Influence Model defensible?** Section 3.11 identifies it

as the framework's weakest link: it is the strongest form of the operant extrapolation, and the extrapolation crosses species, setting, and response class simultaneously. What human evidence exists that unpredictable reward schedules in software produce greater persistence than predictable ones of the same average magnitude, holding value delivered constant? If none exists, the term should be demoted to a hypothesis in the framework file.

1. **Does the taxonomy need a Family I pattern for informational feedback?** Deci et al. report that positive

verbal feedback *enhances* free-choice intrinsic motivation (d = 0.33) where contingent tangible reward undermines it (d = −0.40). Family I currently has no pattern for "tell the user what they achieved in the domain they care about, rather than issuing a token whose value you control." If the family is to be more than a wish-list, this is the strongest evidenced candidate for admission that this chapter found.

1. **Do the load findings scale to the field?** Simons and Chabris (1999) and Rubinstein et al. (2001) are

laboratory results with laboratory tasks. The inference that notification volume imposes a measurable executive-control cost on real users over a working day is marked CITATION NEEDED in section 3.2. It is plausible, it is widely assumed, and it is not, in this document, established.

1. **Is the System 1 / System 2 organizing claim testable?** It is proposed here, not validated. What

experiment would falsify the claim that deceptive patterns work by preventing the recruitment of Type 2 processing rather than by defeating it? A candidate design: hold a pattern constant and vary only the working-memory load imposed elsewhere on the screen. The claim predicts the pattern's effect grows with irrelevant load. As far as this document is aware, that experiment has not been run.

---

## References

Only sources actually cited in this chapter appear here. Every entry was retrieved and read; the URL fetched and the fetch date are recorded in research/sources/04-cognitive-psychology.md for sources first verified for this chapter, and in research/sources/01-dark-pattern-literature.md and

research/sources/02-behavioral-science.md for sources inherited from earlier chapters. Where a claim in this chapter carries a CITATION NEEDED marker, no source is listed, because none was retrieved.

**Dual-process cognition**

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3), 223-241. https://doi.org/10.1177/1745691612460685 *(Record verified against Crossref; SAGE full text returned HTTP 403. Cited for the Type 1 / Type 2 distinction and for the authors' own retreat from "system" to "type." No verbatim quotation is taken from the body.)*

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5), 645-665. https://doi.org/10.1017/S0140525X00003435 *(Abstract retrieved from Cambridge Core; body not retrieved. The abstract contains neither "System 1" nor "System 2." The common attribution of those terms to this paper is marked CITATION NEEDED in section 3.1.)*

**Attention and perception**

Benway, J. P., & Lane, D. M. (1998). Banner Blindness: Web Searchers Often Miss "Obvious" Links. Rice University. https://www.ruf.rice.edu/~lane/papers/banner_blindness.pdf *(Pages 1-2 read. No numeric results are cited from it. The related Human Factors and Ergonomics Society conference version was not retrieved and its pagination is not given.)*

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4), 763-797. https://doi.org/10.1037/0096-1523.27.4.763 *(Full PDF retrieved from apa.org; abstract read directly. No "percentage of time lost to switching" figure is printed, as none appears in the paper.)*

Simons, D. J., & Chabris, C. F. (1999). Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception*, 28(9), 1059-1074. https://doi.org/10.1068/p281059 *(Full PDF retrieved; Method and Results read directly. 192 observers analysed; 54% noticed, 46% did not; Easy task 64%, Hard task 45%.)*

**Memory**

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87-114. https://doi.org/10.1017/S0140525X01003922 *(Full abstract retrieved from Cambridge Core.)*

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63, 81-97. https://psychclassics.yorku.ca/Miller/ *(Cited for Miller's own separation of the two spans and his own scepticism about the number seven. The folk use of "7±2" as a working-memory design rule is not supported by the paper and is not used in this document.)*

**Habit and reinforcement**

Eyal, N., with Hoover, R. (2014). *Hooked: How to Build Habit-Forming Products*. New York: Portfolio / Penguin. ISBN 9781591847786. *(Practitioner framework with no primary empirical base of its own. Cited as such.)*

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of Reinforcement*. New York: Appleton-Century-Crofts. *(Cited at book level. No response-rate or resistance-to-extinction figure is attached, as none was retrieved from the text. The application to human smartphone behavior is an analogy, not a finding in this work.)*

**Motivation**

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627-668. https://doi.org/10.1037/0033-2909.125.6.627 *(Full PDF retrieved; abstract read directly. 128 studies. d = –0.40, –0.36, –0.28 for engagement-, completion- and performance-contingent rewards on free-choice intrinsic motivation; positive feedback d = 0.33.)*

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 28(1), 129-137. https://doi.org/10.1037/h0035519 *(Full PDF retrieved; abstract read directly. Three conditions: expected-award, unexpected-award, no-award. No sample size or completion percentage is printed, as neither was read from the primary text.)*

**Value, loss, and endowment**

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325-1348. https://doi.org/10.1086/261737 *(Full PDF retrieved; abstract read directly. No WTA/WTP ratios are printed, as none were read from the body.)*

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263-291. https://doi.org/10.2307/1914185 *(Verified in research/sources/02-behavioral-science.md. The title is "decision under risk.")*

Ruggeri, K., Alí, S., Berge, M. L., et al. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 4(6), 622-633. https://doi.org/10.1038/s41562-020-0886-x *(4,098 participants, 19 countries, 13 languages; replicated for 94% of items; 12 of 13 contrasts replicated.)*

**Goals, progress, and resumption**

Ghibellini, R., & Meier, B. (2025). Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects. *Humanities and Social Sciences Communications*, 12, Article 962. https://doi.org/10.1057/s41599-025-05000-w *(Pooled recall ratio 0.99, k = 38. Pooled resumption rate 67.0%, k = 21.)*

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *Journal of Marketing Research*, 43(1), 39-58. https://doi.org/10.1509/jmkr.43.1.39 *(Full PDF retrieved; abstract read directly. The paper*

*attributes the original hypothesis to Clark Hull; the Hull primary sources were not retrieved and are cited only via Kivetz et al.)*

Nunes, J. C., & Drèze, X. (2006). The Endowed Progress Effect: How Artificial Advancement Increases Effort. *Journal of Consumer Research*, 32(4), 504-512. https://doi.org/10.1086/500480 *(Cited for the 8-versus-10-step manipulation and the direction of the effect, both confirmed from the primary abstract. The widely circulated completion percentages are not confirmed against the primary text and are not printed in this document.)*

**Social cognition**

Bohner, G., & Schlüter, L. E. (2014). A Room with a Viewpoint Revisited: Descriptive Norms and Hotel Guests' Towel Reuse Behavior. *PLOS ONE*, 9, e104086. https://doi.org/10.1371/journal.pone.0104086 *(Study 1 N = 724; Study 2 N = 204; two German hotels. The descriptive-norm advantage did not replicate.)*

Cialdini, R. B. (2021). *Influence, New and Expanded: The Psychology of Persuasion*. New York: Harper Business. ISBN 9780062937650. *(Seven principles, per https://www.influenceatwork.com/7-principles-of-persuasion/. A different page on the same official site still says six. Replication is mixed and principle-dependent.)*

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research*, 35(3), 472-482. https://doi.org/10.1086/586910 *(Full abstract retrieved from Oxford Academic. No per-condition reuse percentages are printed.)*

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 81. https://doi.org/10.1145/3359183

**Emotion**

Przybylski, A. K., Murayama, K., DeHaan, C. R., & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, 29(4), 1841-1848. https://doi.org/10.1016/j.chb.2013.02.014 *(Bibliographic record verified against Crossref. **The primary text was not retrieved**; the correlational findings summarised in section 3.10 come from secondary descriptions and are marked* CITATION NEEDED *.)*

**Defaults**

Johnson, E. J., & Goldstein, D. (2003). Do Defaults Save Lives? *Science*, 302(5649), 1338-1339. https://doi.org/10.1126/science.1091721 *(Record verified against Crossref; science.org returned HTTP 403. Direction and significance of the opt-in versus opt-out finding cited from a secondary summary (besci.org). Per-country consent percentages were not retrieved and are not printed.)*

**Replication**

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., et al. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, 11(4), 546-573.

https://doi.org/10.1177/1745691616652873 *(Full PDF retrieved; abstract read directly. k = 23 laboratories; N = 2,141; d = 0.04, 95% CI [−0.07, 0.15].)*

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716 *(Abstract retrieved verbatim from PubMed. 100 studies; 97% of originals significant; 36% of replications significant; replication effects half the magnitude of original effects.)*

**Frameworks referenced, not cited as findings**

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The Final Edition*. New Haven: Yale University Press. *(Framework only. No nudge effect size appears anywhere in this document.)*

**Not cited, and deliberately so**

Zeigarnik, B. (1927). Cited nowhere in this chapter as an established effect. See Ghibellini & Meier (2025).

Any source for "decision fatigue" as an explanation of late-flow compliance. The resource model of self-control it depends on did not survive Hagger et al. (2016). Section 3.11 states what this document uses instead.

THE SPINE

# The taxonomy

116 patterns, 9 families, 3 axes. Mechanism and morality are separated, which is why a family of patterns that are *good* can exist at all.

**Status:** v0.1 draft. Pattern IDs are **stable from this point on**, a pattern may be renamed, merged, or downgraded, but its ID is never reused for a different pattern.

**116 patterns across 9 families.**

---

## How this taxonomy differs from existing work

Prior taxonomies of manipulative interface design are organized around a single question: *how is the user being deceived?* That framing is useful for regulators and useless for designers, because it can only ever produce a list of things not to do.

This taxonomy is organized around a different question: *what is the mechanism, and whose interest does this implementation serve?* Mechanism and morality are separated onto different axes. The consequence is that the same mechanism appears once, not twice, and its ethical status becomes a property of the implementation rather than of the pattern.

A streak (**C-01**) is one mechanism. In a language-learning app used by someone who wants to build a daily habit, it is **ethical**. In a social app that sends escalating guilt messages to stop a user leaving, the same mechanism is **manipulative** (**C-14**). A taxonomy that files "streak" permanently under "dark patterns" cannot express this, and so cannot describe how real products actually work.

This also means the taxonomy contains an entire family (**I**) of patterns that are *good*, autonomy-preserving designs that achieve legitimate business goals. No existing dark-pattern taxonomy has this, because a catalogue of harms has nowhere to put a design that helps.

### THE THREE AXES

| AXIS | VALUES |
|---|---|
| **Family** | A–I (the mechanism) |
| **Lifecycle stage** | Discovery, Install, Registration, Activation, Onboarding, Trust, Habit, Routine, Retention, Monetization, Expansion, Exit |
| **Valence** | Ethical / Persuasive / Manipulative (a property of the *implementation*) |

**Severity ceiling** (0–5) is the worst harm the mechanism can do when implemented at its most manipulative. It is a ceiling, not a score: a pattern with ceiling 5 is not necessarily harmful in a given product. It means it *can* be.

---

### Lineage and honesty note

Some pattern names below are inherited from existing literature and industry vocabulary (for example: roach motel, confirmshaming, privacy zuckering, drip pricing, decoy pricing). These are **not original to this work** and must be attributed once their source is retrieved and verified.

Two tags are used:

- ATTR VERIFIED, attribution **verified**. The source was actually retrieved and is recorded in research/sources/. Safe to cite.
- ATTR PENDING, attribution **pending verification**. The name is known to be inherited, but no source has been retrieved yet. Per the Citation Law in PROJECT.md, **no source may be named for these in the manuscript** until it has been fetched and recorded.

Names with neither tag are this work's own naming, or generic descriptive terms in common use.

**Do not promote** ATTR PENDING **to** ATTR VERIFIED **without a retrieved reference.** See the lineage table at the foot of this file for what each verified name is owed to.

---

### Family A, Attention and Capture

*Mechanisms that acquire and hold attention before any decision is made.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| A-01 | Infinite Scroll ATTR PENDING | Content loads continuously, removing the natural stopping cue that a page boundary provides. | Habit | Persuasive → Manipulative | 4 |
| A-02 | Autoplay | The next unit of content begins without a user decision, converting a choice to continue into a choice to stop. | Routine | Persuasive → Manipulative | 4 |
| A-03 | Badge Signal | A numeric or coloured marker indicating unresolved items, exploiting completion tension. | Habit | Ethical → Manipulative | 3 |
| A-04 | Notification Cascade | Volume and timing of push messages tuned to session frequency rather than user need. | Retention | Persuasive → Manipulative | 4 |
| A-05 | Pull-to-Refresh | A manual gesture that triggers an uncertain reward, structurally identical to a lever pull. | Habit | Persuasive → Manipulative | 4 |
| A-06 | Interstitial Interrupt | Full-screen content inserted between the user and their intended destination. | Discovery | Persuasive → Manipulative | 3 |
| A-07 | Unread Debt | Accumulating unresolved items whose count itself becomes the trigger to return. | Routine | Persuasive | 3 |
| A-08 | Zero-Latency Feed | Preloading that removes the pause in which a user might reconsider continuing. | Habit | Ethical → Persuasive | 2 |
| A-09 | Algorithmic Curation | Ordering content by predicted engagement rather than user-stated preference or chronology. | Habit | Ethical → Manipulative | 5 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| A-10 | Ephemeral Content | Content with an expiry, converting optional viewing into time-pressured viewing. | Routine | Persuasive → Manipulative | 3 |
| A-11 | Presence Signal | Displaying that others are active now, creating a synchronous pull to participate. | Habit | Persuasive | 3 |
| A-12 | Sensory Hook | Sound and haptic feedback conditioned to reward delivery. | Habit | Persuasive → Manipulative | 3 |
| A-13 | Temporal Disorientation | Removing clocks, timestamps, and session-length cues so elapsed time is hard to judge. | Routine | Manipulative | 5 |

## Family B, Onboarding and Activation

*Mechanisms acting between first open and first value.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| B-01 | Progressive Disclosure | Complexity revealed in stages to keep perceived difficulty below the abandonment threshold. | Onboarding | Ethical | 1 |
| B-02 | Completion Meter | A visible progress indicator that converts an open-ended task into a closable one. | Onboarding | Ethical → Persuasive | 2 |
| B-03 | Endowed Progress ATTR PENDING | Progress granted before any work is done, so the user begins already partly committed. | Activation | Persuasive | 2 |
| B-04 | Registration Wall | Value withheld until an account is created, before the value has been demonstrated. | Registration | Persuasive → Manipulative | 3 |
| B-05 | Social Login Shortcut | Reduced signup friction in exchange for identity and graph data the user does not evaluate. | Registration | Ethical → Manipulative | 3 |
| B-06 | Personalization Interview | Questions framed as customisation that also perform commitment and data capture. | Onboarding | Ethical → Persuasive | 2 |
| B-07 | Aha-Moment Engineering | Deliberate sequencing of first-session actions toward the moment predictive of retention. | Activation | Ethical → Persuasive | 2 |
| B-08 | Empty-State Coaching | Using the zero-data state to direct the next action rather than present an inert screen. | Activation | Ethical | 1 |
| B-09 | Permission Priming | A soft in-app request preceding the OS permission dialog, protecting the one-shot system prompt. | Onboarding | Persuasive → Manipulative | 3 |
| B-10 | Commitment Escalation | A trivially small first ask that makes the subsequent larger ask harder to refuse. | Onboarding | Persuasive → Manipulative | 3 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| B-11 | Contact Ingestion Prompt | Requesting the address book at the moment of maximum goodwill and minimum scrutiny. | Onboarding | Persuasive → Manipulative | 4 |

### Family C, Habit and Reinforcement

*Mechanisms that convert a decision into a routine, and a routine into an automatic behavior.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| C-01 | Streak | A count of consecutive periods of use, whose value derives entirely from its fragility. | Habit | Ethical → Manipulative | 4 |
| C-02 | Variable Reward ATTR PENDING | Reward delivered on an unpredictable schedule, producing more persistent behavior than a reliable one. | Habit | Persuasive → Manipulative | 5 |
| C-03 | Login Bonus | A reward contingent on presence rather than on any useful action. | Routine | Persuasive → Manipulative | 3 |
| C-04 | Streak Insurance | A purchasable or earnable repair for a broken streak, monetising the anxiety the streak created. | Monetization | Persuasive → Manipulative | 4 |
| C-05 | Goal Gradient | Effort intensifies as a visible goal nears; interfaces exploit this by making the end visible. | Habit | Ethical → Persuasive | 2 |
| C-06 | Trigger Scheduling | Re-engagement timed to the user's own historical usage windows. | Retention | Persuasive → Manipulative | 3 |
| C-07 | Investment Loop ATTR PENDING | Each use deposits user work into the product, raising the cost of ever leaving. | Retention | Ethical → Manipulative | 4 |
| C-08 | Near Miss | A near-success presented as more motivating than a clear failure, sustaining repetition. | Habit | Manipulative | 5 |
| C-09 | Achievement System | Discrete rewards for defined behaviors, converting use into collection. | Habit | Ethical → Persuasive | 2 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| C-10 | Tier Progression | Escalating status levels whose loss on inactivity creates retention pressure. | Retention | Persuasive → Manipulative | 3 |
| C-11 | Ritual Anchoring | Binding product use to an existing daily routine so an external cue does the triggering. | Habit | Ethical → Persuasive | 2 |
| C-12 | Reward Escalation | Reward magnitude increasing with consecutive engagement, so lapsing forfeits accrued value. | Retention | Persuasive → Manipulative | 4 |
| C-13 | Completion Tension | An unfinished state deliberately left visible, exploiting the pull of the incomplete. | Habit | Persuasive | 3 |
| C-14 | Guilt Messaging | Copy that frames a lapse as a personal or relational failure to compel return. | Retention | Manipulative | 5 |

### Family D, Social and Identity

*Mechanisms that recruit other people, and the user's self-image, as the engine of engagement.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| D-01 | Social Proof Counter ATTR PENDING | Displaying others' behavior as evidence for what the user should do. | Trust | Ethical → Manipulative | 4 |
| D-02 | Reciprocity Prompt | A prompt engineered so an unreturned social action feels like an outstanding debt. | Habit | Persuasive → Manipulative | 3 |
| D-03 | Read Receipt | Making one party's attention visible to the other, converting a message into an obligation. | Routine | Ethical → Manipulative | 4 |
| D-04 | Typing Indicator | Real-time evidence of an imminent reply, holding the user in the interface. | Routine | Persuasive | 2 |
| D-05 | Public Metrics | Publishing counts attached to a person, tying self-worth to a number the product controls. | Habit | Persuasive → Manipulative | 5 |
| D-06 | Comparative Ranking | Ordering users against each other, converting use into competition. | Habit | Persuasive → Manipulative | 4 |
| D-07 | Identity Investment | Accumulated self-presentation that the user would lose by leaving. | Retention | Ethical → Manipulative | 4 |
| D-08 | Parasocial Hook | Engineering one-sided attachment to a creator or persona as the retention mechanism. | Retention | Persuasive → Manipulative | 4 |
| D-09 | Missing-Out Signal | Surfacing what the user did not see, framing absence as loss. | Retention | Persuasive → Manipulative | 4 |
| D-10 | Status Marker | A conferred visual distinction whose scarcity is set by the platform. | Monetization | Ethical → Persuasive | 2 |
| D-11 | Group Obligation | Shared goals or streaks so that lapsing harms other people, not only oneself. | Retention | Persuasive → Manipulative | 4 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| D-12 | Reputation Lock-in | Earned standing that cannot be exported, so leaving means starting from zero. | Exit | Persuasive → Manipulative | 4 |

## Family E, Growth and Acquisition

*Mechanisms that use existing users to obtain new ones.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| E-01 | Referral Incentive | A reward for bringing another user, aligning user and business interest when the product is good. | Expansion | Ethical → Persuasive | 2 |
| E-02 | Viral Invite Loop | An invite step embedded in the core flow so growth is a by-product of normal use. | Expansion | Ethical → Manipulative | 3 |
| E-03 | Contact Harvesting | Uploading the address book, converting the user's relationships into the company's asset. | Expansion | Persuasive → Manipulative | 5 |
| E-04 | Invite Impersonation | Outbound messages framed as if from the user, sent with consent that was not meaningfully given. | Expansion | Manipulative | 5 |
| E-05 | Share-to-Unlock | Value gated behind a public share the user would not otherwise make. | Expansion | Persuasive → Manipulative | 3 |
| E-06 | Watermarked Artifact | Product output carrying an advertisement the user distributes on the company's behalf. | Expansion | Ethical → Persuasive | 2 |
| E-07 | Network Pressure | Surfacing that known contacts are present, converting a product choice into a social one. | Install | Persuasive → Manipulative | 3 |
| E-08 | Waitlist Scarcity | Artificial queueing to manufacture the perception of demand. | Discovery | Persuasive → Manipulative | 3 |
| E-09 | Exclusivity Gate | Invite-only access, using restriction itself as the attractor. | Discovery | Persuasive | 2 |
| E-10 | Programmatic Surface Farming | Generating landing pages at scale to intercept intent | Discovery | Ethical → Manipulative | 3 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| | | before a competitor does. | | | |
| E-11 | Deferred Continuity | Preserving intent across the install boundary so context is not lost at the app-store gap. | Install | Ethical | 1 |
| E-12 | Install Interstitial | Obstructing the mobile web experience to force an app install. | Install | Persuasive → Manipulative | 3 |

## Family F, Monetization and Pricing

*Mechanisms that convert behavior into revenue.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| F-01 | Capability Wall | A feature boundary between free and paid, chosen to be felt at the moment of highest need. | Monetization | Ethical → Manipulative | 3 |
| F-02 | Metered Paywall | Access limited by quantity, so the wall arrives once the habit exists. | Monetization | Ethical → Persuasive | 2 |
| F-03 | Trial-to-Paid Conversion | A free period ending in a charge, whose ethics depend entirely on the salience of that ending. | Monetization | Ethical → Manipulative | 4 |
| F-04 | Drip Pricing ATTR PENDING | Cost revealed in increments so the true total appears only at the final step. | Monetization | Manipulative | 5 |
| F-05 | Price Anchoring | A reference price presented to make the target price seem small by comparison. | Monetization | Persuasive → Manipulative | 3 |
| F-06 | Decoy Option ATTR PENDING | An option included not to be chosen but to make another option look better. | Monetization | Persuasive → Manipulative | 3 |
| F-07 | Charm Pricing ATTR PENDING | Prices set just below a round boundary to exploit left-digit processing. | Monetization | Persuasive | 2 |
| F-08 | Confirmshaming ATTR VERIFIED | The decline option worded to shame the user for declining. | Monetization | Manipulative | 4 |
| F-09 | Roach Motel ATTR VERIFIED | Entry made trivially easy and exit made deliberately hard. | Exit | Manipulative | 5 |
| F-10 | Forced Continuity ATTR VERIFIED | Automatic charging at trial end with the renewal deliberately de-emphasised. | Monetization | Manipulative | 5 |
| F-11 | Concealed Subscription | A recurring charge that the user did not understand to be recurring. | Monetization | Manipulative | 5 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| F-12 | Preselected Add-on | An extra opted in by default, relying on the user not to notice. | Monetization | Manipulative | 4 |
| F-13 | Basket Sneaking ATTR VERIFIED | An item the user did not select appearing in the cart. | Monetization | Manipulative | 5 |
| F-14 | Manufactured Urgency | A countdown or deadline that is not real, or that silently resets. | Monetization | Manipulative | 5 |
| F-15 | Scarcity Claim | A stock or availability claim presented as fact without being verifiable. | Monetization | Persuasive → Manipulative | 5 |
| F-16 | Randomized Purchase | Paying for an uncertain outcome, structurally a wager. | Monetization | Manipulative | 5 |
| F-17 | Currency Obfuscation | An intermediate token that decouples spending from the perception of money. | Monetization | Persuasive → Manipulative | 5 |
| F-18 | Comparison Obstruction ATTR VERIFIED | Design that makes it hard to compare options or to see the unit price. | Monetization | Manipulative | 4 |

## Family G, Consent, Data and Privacy

*Mechanisms acting on the user's ability to understand and control what is taken from them.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| G-01 | Consent Asymmetry | Accept made one click and decline made many, so the cheap path is the permissive one. | Trust | Manipulative | 5 |
| G-02 | Preselected Opt-in | Consent granted by default, so inaction is read as agreement. | Trust | Manipulative | 5 |
| G-03 | Default Over-disclosure ATTR VERIFIED | Default settings that share more than a reasonable user would choose. | Trust | Manipulative | 5 |
| G-04 | Bundled Consent | Multiple unrelated permissions collapsed into a single accept. | Trust | Manipulative | 5 |
| G-05 | Permission Nagging ATTR VERIFIED | Repeating a rejected request until the user yields. | Trust | Manipulative | 4 |
| G-06 | Setting Obscurity | Controls that exist, satisfying the letter of the law, but are placed where they will not be found. | Trust | Manipulative | 4 |
| G-07 | Default Publicity | New content or profiles public unless the user acts, with the default unstated. | Trust | Manipulative | 5 |
| G-08 | Consent Nudge | Visual weight, colour, and copy steering the user toward the permissive option. | Trust | Manipulative | 4 |
| G-09 | Export Obstruction | A data-export right honoured in a form too degraded to be useful. | Exit | Manipulative | 4 |
| G-10 | Inferred Profile | Building a profile from data the user never knowingly provided. | Trust | Manipulative | 5 |
| G-11 | Consent Fatigue | Exhausting the user's scrutiny with volume until they stop reading. | Trust | Manipulative | 4 |
| G-12 | Ambiguous Wording | Double negatives and inverted phrasing so the user cannot tell | Trust | Manipulative | 5 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| | | what accepting means. | | | |

### Family H, Retention and Exit Friction

*Mechanisms that act on the user who has decided to leave.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| H-01 | Cancellation Maze | Steps, screens, and detours inserted between the intention to cancel and cancelling. | Exit | Manipulative | 5 |
| H-02 | Exit Offer | A discount presented at cancellation, which is honest retention or a delay tactic depending on whether cancelling remains one click away. | Exit | Ethical → Manipulative | 3 |
| H-03 | Guilt-Framed Exit | Cancellation copy engineered to make leaving feel like a moral failure. | Exit | Manipulative | 4 |
| H-04 | Deletion Obstruction | Account deletion made materially harder than account creation. | Exit | Manipulative | 5 |
| H-05 | Reactivation Window | A delay before deletion takes effect, during which the user is lobbied to return. | Exit | Persuasive → Manipulative | 3 |
| H-06 | Data Hostage | Content the user created, withheld or degraded on departure. | Exit | Manipulative | 5 |
| H-07 | Downgrade Penalty | Punitive loss of unrelated function on downgrade, deterring any reduction in spend. | Exit | Persuasive → Manipulative | 4 |
| H-08 | Win-back Blast | High-volume outreach to a user who has already signalled departure. | Exit | Persuasive → Manipulative | 3 |
| H-09 | Dormancy Nag | Escalating contact with a user whose non-use is itself the signal being ignored. | Retention | Persuasive → Manipulative | 3 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE RANGE | CEILING |
|---|---|---|---|---|---|
| H-10 | Pause as Trap | A pause offered in place of cancellation, keeping the account and its billing relationship alive. | Exit | Persuasive → Manipulative | 4 |
| H-11 | Support Obstruction | Human help routed away from the user precisely when the request would cost the company money. | Exit | Manipulative | 4 |
| H-12 | Notification Escalation | Increasing message volume as engagement falls, accelerating the disengagement it means to reverse. | Retention | Manipulative | 3 |

### Family I, Trust and Autonomy Preservation

*Mechanisms that achieve legitimate business goals without extracting from the user. This family has no counterpart in existing dark-pattern taxonomies. It is where the document's normative argument is actually made: not by condemning patterns, but by demonstrating that the business goal each harmful pattern serves has a non-harmful solution.*

| ID | PATTERN | DEFINITION | STAGE | VALENCE | NOTES |
|---|---|---|---|---|---|
| I-01 | Honest Default | The default set to what a reasonable user would choose if they were paying full attention. | Trust | Ethical | Counterpart to G-02, G-07 |
| I-02 | Symmetric Consent | Accept and decline given equal visual weight and equal cost in clicks. | Trust | Ethical | Counterpart to G-01, G-08 |
| I-03 | Frictionless Exit | Cancellation no harder than signup, in the same place, in the same number of steps. | Exit | Ethical | Counterpart to F-09, H-01 |
| I-04 | Usage Transparency | Showing the user their own consumption honestly, including when it is unflattering. | Routine | Ethical | Counterpart to A-13 |
| I-05 | Sufficiency Nudge | Actively suggesting the user has had enough for now. | Routine | Ethical | Counterpart to A-01, A-02 |
| I-06 | Stopping Cue | A designed end to a unit of content, restoring the boundary infinite scroll removed. | Habit | Ethical | Counterpart to A-01 |
| I-07 | Undo Over Confirm | Allowing the action and making it reversible, rather than obstructing it with a dialog. | Routine | Ethical | Reduces cost on every axis |
| I-08 | Reversible Commitment | Any commitment can be unwound on the same terms it was made. | Monetization | Ethical | Counterpart to F-09, F-10 |
| I-09 | Verifiable Scarcity | Scarcity claims stated only when true and checkable. | Monetization | Ethical | Counterpart to F-14, F-15 |
| I-10 | Unused-Value Alert | Proactively telling a user they are paying for something they do not use. | Monetization | Ethical | Counterpart to F-11, H-09 |
| I-11 | Portability by Design | Export built as a first-class feature, not a compliance obligation. | Exit | Ethical | Counterpart to G-09, H-06, D-12 |

| ID | PATTERN | DEFINITION | STAGE | VALENCE | NOTES |
|---|---|---|---|---|---|
| I-12 | Consumption Budget | User-set limits the product enforces against its own engagement interest. | Routine | Ethical | Counterpart to A-04, C-01 |

## Counts

| FAMILY | PATTERNS |
|---|---|
| A, Attention and Capture | 13 |
| B, Onboarding and Activation | 11 |
| C, Habit and Reinforcement | 14 |
| D, Social and Identity | 12 |
| E, Growth and Acquisition | 12 |
| F, Monetization and Pricing | 18 |
| G, Consent, Data and Privacy | 12 |
| H, Retention and Exit Friction | 12 |
| I, Trust and Autonomy Preservation | 12 |
| **Total** | **116** |

## Position against the existing taxonomies

All figures below are **verified** against retrieved sources (research/sources/01-dark-pattern-literature.md).

| TAXONOMY | TYPES | ORGANIZING QUESTION | COVERS NON-HARMFUL PATTERNS? |
|---|---|---|---|
| Brignull, deceptive.design (2010–) | 18 | How is the user deceived? | No |
| Gray et al., CHI 2018 | 5 strategies | What is the designer's strategy? | No |
| Mathur et al., CSCW 2019 | 15 types / 7 categories | What is empirically observable at scale? | No |
| Gray et al., CHI 2024 ontology | 64 types, 3 levels | How do we harmonise 10 taxonomies? | No |
| **This work** | **116, 9 families, 3 axes** | **What is the mechanism, and whose interest does this implementation serve?** | **Yes, Family I** |

The count is larger not because we found more ways to harm users, but because **we admit patterns that are not harms.** Every taxonomy above is a catalogue of deception, so each one structurally excludes the ethical and persuasive majority of behavioral design. Ours does not.

The honest corollary: our 116 and their 64 are **not comparable numbers** and must never be presented as though the larger one is better. Gray et al. (2024) harmonised ten taxonomies to converge on a shared regulatory vocabulary, a different and harder job than ours. Where our names overlap theirs, we defer to theirs.

### Attribution lineage

Names inherited from prior work, with the source now verified and retrieved.

| ID | NAME AS USED HERE | OWED TO | STATUS |
|---|---|---|---|
| F-08 | Confirmshaming | Brignull, deceptive.design /types | ✓ verified |
| F-09 | Roach Motel | Brignull (reproduced in Gray et al. 2018, Table 1) | ✓ verified |
| F-10 | Forced Continuity | Brignull (Gray et al. 2018, Table 1) | ✓ verified |
| F-13 | Basket Sneaking (orig. "Sneak into Basket") | Brignull; measured by Mathur et al. 2019 | ✓ verified |
| F-18 | Comparison Obstruction (orig. "Price Comparison Prevention") | Brignull (Gray et al. 2018, Table 1) | ✓ verified |
| G-03 | Default Over-disclosure (orig. "Privacy Zuckering") | Brignull (Gray et al. 2018, Table 1) | ✓ verified |
| G-05 | Permission Nagging (orig. "Nagging") | Gray et al. 2018, strategy category | ✓ verified |
| B-03 | Endowed Progress | Nunes & Drèze, *J. Consumer Research* 2006, DOI 10.1086/500480 | ✓ verified |
| C-02 | Variable Reward | **Ferster & Skinner (1957)**, *Schedules of Reinforcement*, note it is **not** "Skinner (1957)"; omitting Ferster is the single most common citation error in this literature | ✓ verified |
| C-07 | Investment Loop | Eyal, *Hooked* (2014), Portfolio/Penguin, practitioner framework, **no empirical base of its own** | ✓ verified |
| D-01 | Social Proof | Cialdini, *Influence* (2021 expanded ed.), now **seven** principles, not six; Unity was added | ✓ verified |
| C-13 | Completion Tension | **Ovsiankina effect**, *not* Zeigarnik. See correction below. | ✓ verified |
| A-01 | Infinite Scroll | Attribution to Aza Raskin is widely repeated but **not retrieved**. | ? pending |
| F-04 | Drip Pricing | Economics/regulatory term, **not yet retrieved** | ? pending |
| F-06 | Decoy Option | Asymmetric dominance literature, **not yet retrieved** | ? pending |
| F-07 | Charm Pricing | **not yet retrieved** | ? pending |

### Two corrections that must propagate into the manuscript

These were caught by source verification and would otherwise have shipped as confident errors. Both are recorded here because both were present in the project's own opening plan.

#### 1. "BEHAVIOR = MOTIVATION × ABILITY × TRIGGER" IS NOT FOGG'S EQUATION

The formula B=MAT **does not appear anywhere in Fogg (2009)**. The full 7-page paper was retrieved and read; it contains **no equation at all**. The model is stated only in prose and in two figures. The mnemonic is a later retrofit by the product community, and the multiplicative form in particular implies a mathematical claim Fogg never made.

Further: Fogg **renamed "Trigger" to "Prompt" in late 2017** (confirmed verbatim on his own site). The current formulation is **B=MAP**.

Consequences for this document:

- Never write "Fogg's equation B = MAT." Write: *Fogg (2009) proposes that behavior is a product

of three factors, motivation, ability, and triggers* (his word in 2009; now "prompts").

- Our **Behavioral Influence Model** is our own construct. It is *inspired by* Fogg, not derived

from him, and it must say so. Presenting it as an extension of "Fogg's equation" would be extending something that does not exist.

#### 2. THE ZEIGARNIK EFFECT DOES NOT REPLICATE

C-13 (Completion Tension) was originally conceived on the Zeigarnik effect, the claim that interrupted tasks are better *remembered*. A 2025 meta-analysis (Ghibellini & Meier, *Humanities and Social Sciences Communications* 12:962, DOI 10.1057/s41599-025-05000-w) finds the memory effect essentially absent: pooled recall ratio **0.99** for interrupted vs completed tasks. Their conclusion: the Zeigarnik effect "lacks universal validity."

What *does* replicate is the **Ovsiankina effect**, the tendency to **resume** an interrupted task. Pooled resumption rate **67.0%**.

This is good news for the taxonomy, not bad. Progress bars, incomplete profiles, and unfinished onboarding are **resumption** phenomena, not memory phenomena. The correct mechanism was always Ovsiankina. C-13 is therefore grounded in a *stronger* effect than the one product literature habitually cites, and the document can make that point, which is itself a small original contribution, since citing "Zeigarnik" for this is near-universal in UX writing and is wrong.

**Rule going forward:** any pattern resting on a contested effect must state its replication status in the entry. Rigor here is a feature. A report that says "this well-known effect does not replicate, here is the one that does" is a stronger report than one that repeats the folklore.

Note on renaming: where we have renamed an inherited pattern (F-18, G-03), the original name is recorded above. "Privacy Zuckering" is a proper noun aimed at an individual; the mechanism is better described neutrally, and a document arguing for rigor should not carry a joke in its index. The original is preserved for citation, not for use.

### Terminology

We use **"deceptive patterns"** rather than "dark patterns" in the manuscript. This is not euphemism. Brignull, who coined the term in 2010, now explicitly prefers "manipulative, deceptive and coercive

patterns" and renamed his own site from darkpatterns.org to deceptive.design on the advice of the Tech Policy Design Lab, on the grounds that "dark" carries harmful connotations. When the originator of a term abandons it, continuing to use it is a choice that needs defending. We follow him. *(Verified: deceptive.design/about-us.)*

The term "dark patterns" is retained only when quoting or citing sources that use it, including the statutes that codify it.

---

### Open questions on the taxonomy itself

These are real unresolved problems, recorded rather than hidden.

1. **Is A-09 (Algorithmic Curation) one pattern or a family?** It plausibly deserves its own

family in the AI chapter. Currently under-modelled.

1. **F-16 and F-17 may belong to a gambling sub-family** rather than to monetization, since

their regulatory treatment differs in kind, not degree.

1. **Family I risks becoming a wish-list.** Each entry needs a *shipped, named product* that

actually does it, or it should be cut. This is the hardest evidentiary burden in the document.

1. **Severity ceilings are currently assigned by judgment, not by instrument.** They must be

re-derived from the Behavioral Cost Index once that framework is formalised, or they are just opinions with numbers attached.

1. **Overlap between C-07, D-07, and D-12** (investment, identity, reputation) is unresolved.

All three describe accumulated switching cost by different substrates. They may collapse.

ORIGINAL CONTRIBUTIONS

# The frameworks

Six instruments proposed here and labelled as such. None is dressed as an established finding.

## The Behavioral Influence Model (BIM)

**Status: original framework; proposed here.**

### PROVENANCE, STATED HONESTLY

This model is **inspired by** B.J. Fogg's Behavior Model (2009). It is **not derived from it**, and it is not an extension of any equation of his.

This needs saying plainly, because the product literature routinely attributes an equation to Fogg that he never wrote. The 2009 paper was retrieved and read in full for this project. **It contains no equation.** Fogg states his model in prose and in two figures: behavior is a product of three factors, motivation, ability, and triggers. The formula B = MAT is a community retrofit. Its multiplicative form asserts a mathematical relationship Fogg never claimed. Fogg also renamed "trigger" to "prompt" in late 2017, so even the mnemonic's final letter is out of date.

We therefore make our own claim, in our own name, and take responsibility for it.

### THE MODEL

Fogg's three factors describe whether a behavior *fires* at a moment of opportunity. They do not describe whether a behavior *persists*, and they say nothing about what the behavior costs the person performing it. A model of consumer software needs both, because consumer software is not in the business of causing one action. It is in the business of causing the same action ten thousand times.

We propose:

$$\text{Behavior Persistence} = \frac{(\text{Motivation} \times \text{Ability} \times \text{Trigger}) \times \text{Reward Uncertainty} \times \text{Repetition}}{\text{Friction} \times \text{Perceived Cost}}$$

Read as: the numerator is what a product adds; the denominator is what the user pays. Design work is the manipulation of this ratio, and the ethics of design work is a question about which term you chose to move.

**The terms**

| TERM | DEFINITION | WHO CONTROLS IT |
|---|---|---|
| **Motivation** | The user's want, at the moment of the prompt. | Mostly the user |
| **Ability** | How easy the action is to perform. | The product |
| **Trigger** | The cue that occasions the action. | The product |
| **Reward Uncertainty** | The variance of the payoff, not its size. | The product |
| **Repetition** | Prior instances of the same loop. | Accumulates |
| **Friction** | Effort standing between intent and action. | The product |
| **Perceived Cost** | What the user *believes* the action costs them. | The product |

---

### THE FOUR CLAIMS THIS MODEL MAKES THAT FOGG'S DOES NOT

**1. Uncertainty is a multiplier, not a reward size.** A reliable reward and an unpredictable reward of the same average magnitude do not produce the same persistence. This is the oldest finding in operant conditioning (Ferster & Skinner, 1957), and it is why the term is *Reward Uncertainty* rather than *Reward*. A product that makes its payoff less predictable has increased persistence **without increasing value delivered.** That sentence is the whole ethical problem of consumer software in miniature.

**2. Perceived cost, not actual cost, sits in the denominator.** This is the load-bearing move. A product can increase persistence by *reducing the cost to the user*, or by *reducing the user's perception of a cost it has not actually reduced.* Both raise the ratio identically. The equation cannot tell them apart, and neither, from the inside, can the metrics dashboard.

This is where deceptive patterns come from. They are not a separate phenomenon requiring a separate theory. **They are the cheapest available term to move.** Currency obfuscation (F-17) lowers perceived cost while actual cost is unchanged. Drip pricing (F-04) does the same to money. Temporal disorientation (A-13) does it to time.

**3. Friction is signed.** Fogg's model treats difficulty as something to remove. But a product removes friction on the path it wants (signup) and *adds* friction on the path it does not (cancellation). Friction is

not a property of a product; it is a **field with a direction**, and its asymmetry is a measurement of intent. See the Behavioral Gravity Model.

The single most diagnostic question you can ask of any consumer product is: *how many clicks to start, how many to stop?* Roach Motel (F-09) is precisely the deliberate divergence of these two numbers.

**4. Repetition is in the numerator, so the model is recursive.** Each firing of the loop raises the probability of the next. Habit is not a separate stage that follows behavior; it is behavior with a high repetition term. This means the model predicts its own runaway, and explains why habit-forming products exhibit hockey-stick retention curves rather than linear ones.

---

#### WHAT THE MODEL IS FOR

It is a **diagnostic instrument**, not a prediction engine. The terms are not measured in comparable units and the equation should not be computed. Anyone who reports a BIM "score" has misunderstood it.

Its use is to force a specific question about any interface decision:

### Which term did this change move, and did it move it by delivering value or by distorting perception?

That question sorts behavioral design into the three valences.

| THE CHANGE... | VALENCE |
|---|---|
| raises Motivation by making the product better | **Ethical** |
| raises Ability or lowers Friction on a path the user wants | **Ethical** |
| raises Trigger quality by timing it to genuine need | **Ethical** |
| raises Reward Uncertainty to increase persistence | **Persuasive**, tipping to Manipulative |
| lowers *Perceived* Cost without lowering actual cost | **Manipulative**, always |
| raises Friction on the exit path | **Manipulative**, always |

The last two rows are the definition of a deceptive pattern, derived rather than asserted. This is the argument the taxonomy makes structurally, and the reason Family I exists: every row in the manipulative half has an ethical counterpart that reaches the same business metric by moving a different term.

---

#### WORKED EXAMPLE: THE STREAK (C-01)

| TERM | WHAT THE STREAK DOES |
|---|---|
| Motivation | Adds an extrinsic goal that did not previously exist |
| Ability | Unchanged |
| Trigger | Supplies a daily one, and a reason to obey it |
| Reward Uncertainty | Low. The streak reward is perfectly predictable |
| Repetition | High, by construction. The streak *is* a repetition counter |
| Friction | Unchanged |
| Perceived Cost | **This is where the ethics live** |

A streak in a language app: the user *wants* the daily habit. Motivation is aligned; perceived cost is honest; the extrinsic goal scaffolds an intrinsic one. **Ethical.**

The same streak, with escalating guilt messaging on lapse (C-14), and a purchasable repair for the broken streak (C-04): the product has manufactured an anxiety and then sold the cure. The perceived cost of *stopping* has been inflated far above its real cost, which is zero, nothing is actually lost by missing a day of an app. **Manipulative.**

Identical mechanism. Different term moved. That is the entire thesis of this document, and the BIM is what makes it a derivation instead of an opinion.

---

#### LIMITATIONS

1. The terms are not independent. Repetition raises Motivation; Friction shapes Perceived Cost.

The multiplicative form implies an independence that does not hold. **This is the same criticism we levelled at B = MAT, and intellectual honesty requires levelling it at ourselves.** The model is a lens, not an equation, and the notation is a convenience that should not be mistaken for a claim about arithmetic.

1. It says nothing about *which* behavior. It models the strength of a loop, not the value of

what the loop produces.

1. It is not empirically validated. It is a synthesis, offered for use and for refutation.

#### RELATED

Behavioral Gravity Model (friction as a signed field) · Behavioral Cost Index (unpacks Perceived Cost into seven axes) · Behavioral Debt (what accumulates when Perceived Cost and actual cost diverge for long enough)

## The Behavioral Cost Index (BCI)

**Status: original framework; proposed here.**

---

### THE PROBLEM IT SOLVES

Every existing taxonomy of deceptive design classifies patterns by **what the designer did**. None of them measure **what the user paid**.

This is a real gap, and it has a practical consequence: severity ratings in the existing literature are assigned by judgment. A pattern is called "serious" because it feels serious. That is not an instrument, and it cannot survive a hostile question, *why is this a 4 and not a 3?*

The Behavioral Cost Index is an attempt at an answer. It rests on one observation:

**Every interaction imposes a cost on the user, and the cost is never only time.**

A cookie banner costs almost no time and a great deal of privacy. A cancellation maze costs time and trust and almost nothing else. A loot box costs money and emotion. These are not the same harm, and collapsing them onto a single "severity" number destroys the information that makes them different.

---

### THE SEVEN AXES

| AXIS | WHAT IS SPENT | DIAGNOSTIC QUESTION |
|---|---|---|
| **Time** | Minutes of life | Would the user have chosen to spend this long, if asked in advance? |
| **Attention** | Cognitive bandwidth | Was the user's focus taken, or given? |
| **Memory** | Working memory load | Must the user hold something in mind to avoid being caught out? |
| **Trust** | Willingness to believe the product next time | Would the user feel betrayed on learning what happened? |
| **Money** | Currency | Did the user pay more than they believed they were paying? |
| **Privacy** | Information about the self | Did the user disclose more than they understood they were disclosing? |
| **Emotion** | Affective state | Did the product make the user feel bad on purpose? |

Each is scored **0 to 5**.

**The scoring rule (this is what makes it an instrument rather than a vibe):**

| SCORE | MEANING |
|---|---|
| 0 | No cost on this axis |
| 1 | Cost imposed, and the user would have accepted it if asked plainly |
| 2 | Cost imposed, user is aware, mild friction |
| 3 | Cost imposed, and the user is *not fully aware* of it |
| 4 | Cost imposed, user is unaware, and awareness would change their behavior |
| 5 | Cost imposed, user is unaware, awareness would change their behavior, **and the design works precisely because of that unawareness** |

Note what the scale is actually measuring. It is not measuring magnitude. It is measuring the **gap between the cost paid and the cost understood.**

That gap is the definition of manipulation.

---

### THE DERIVATION OF VALENCE

This is the payoff, and it is why the BCI matters more than a severity number.

Valence is not assigned. It **falls out of the scores**:

- **Ethical**, every axis scores ≤ 2. Costs exist, but the user sees them.
- **Persuasive**, highest axis is 3. The user is being steered, but the steering is

survivable under scrutiny.

- **Manipulative**, any axis scores 4 or 5. The design depends on the user not knowing.

A single 5 on any axis is sufficient for manipulation. Harms do not average out. A product that is exemplary on six axes and scores 5 on Privacy is not "mostly ethical."

**Severity ceiling** = the highest score the pattern can reach on its worst axis, in its worst implementation. This replaces the judgment calls currently sitting in TAXONOMY.md, which are flagged there as an open problem precisely because they were assigned by intuition.

---

### WORKED EXAMPLES

**G-01, Consent Asymmetry (accept in one click, decline in six)**

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 2 | The extra clicks are real but small |
| Attention | 3 | Requires vigilance the user did not budget for |
| Memory | 1 | Little to hold in mind |
| Trust | 4 | The asymmetry is legible once seen, and reads as contempt |
| Money | 0 | None |
| **Privacy** | **5** | **The design's entire function is to obtain consent that would not be given under symmetric conditions** |
| Emotion | 2 | Irritation |

**Max = 5 → Manipulative.** Not because we disapprove of it, but because it fails the test: it works *because* the user does not do the arithmetic.

#### I-03, Frictionless Exit (cancel in the same place, same clicks, as signup)

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 0 | Costs the user nothing beyond the action itself |
| Attention | 0 | Nothing to notice |
| Memory | 0 | Nothing to retain |
| Trust | 0 | Builds it rather than spending it |
| Money | 0 | None |
| Privacy | 0 | None |
| Emotion | 0 | None |

**Max = 0 → Ethical.** Note that this pattern still *serves a business goal*, a user who trusts that they can leave is more willing to arrive. Family I is not charity.

#### C-01, Streak, the same pattern, scored twice

This is the case that proves the instrument, because the pattern is identical and only the implementation differs.

| AXIS | LANGUAGE APP | WITH GUILT MESSAGING + PAID REPAIR |
|---|---|---|
| Time | 1 | 2 |
| Attention | 1 | 3 |
| Memory | 1 | 1 |
| Trust | 0 | 3 |
| Money | 0 | **4** |
| Privacy | 0 | 0 |
| **Emotion** | 1 | **5** |

**Left: max 1 → Ethical. Right: max 5 → Manipulative.**

The Emotion 5 is the crux. The product manufactures an anxiety that has no referent in reality, nothing is genuinely lost by missing a day, and then sells the remedy for it (C-04). It works *because* the user has not noticed that the thing they are afraid of losing was created by the company that is now charging them to protect it.

A taxonomy that files "streak" under dark patterns cannot say this. A taxonomy that files it under gamification cannot say it either. The BCI can say it, and can show its working.

---

### AGGREGATE MEASURES

For an app teardown, two summary statistics:

**Behavioral Cost Load** = sum of the maximum score on each axis across all patterns present. Range 0 to 35. Answers: *how expensive is this product to use?*

**Opacity** = the count of axes on which the product scores 4 or 5. Range 0 to 7. Answers: *how much of that cost is hidden?*

Opacity is the more important number, and it is the one no existing framework reports. A product can be expensive and honest. A product that is cheap and opaque is worse, because the user cannot consent to a cost they cannot see.

---

### LIMITATIONS

1. **Scores are still assigned by a human reader.** The BCI does not remove judgment; it

*localises* it. The judgment now attaches to a specific, arguable question ("would awareness change their behavior?") rather than to a vague one ("how bad is this?"). That is an improvement, not a solution, and it should not be oversold.

1. **Inter-rater reliability is untested.** Two analysts may score the same pattern differently.

Establishing that they do not would require an actual study, which this document has not run. This is stated as a limitation, not hidden.

1. **The axes are not independent.** Trust is partly a function of the other six.
2. **A 0-5 ordinal scale is not an interval scale.** Summing it, as Behavioral Cost Load does, is

formally dubious. It is offered as a comparative heuristic, not a measurement.

RELATED

Behavioral Influence Model (the BCI unpacks its *Perceived Cost* term) · Behavioral Debt (what accrues when Opacity stays high over time) · TAXONOMY.md severity ceilings, which this framework is intended to replace

---

## Four Supporting Frameworks

**Status: all original; proposed here.**

Behavioral Pattern DNA · The Behavioral Genome · Behavioral Gravity · Behavioral Debt

---

### 1. Behavioral Pattern DNA

The universal description schema. Every pattern in this document decomposes into the same nine links, and the chain is what makes 116 heterogeneous patterns comparable to one another.

The chain is a **loop**, not a line. The final link funds the first. This is why deceptive patterns proliferate without anyone deciding to be malicious: a pattern that moves a metric justifies further investment in patterns that move that metric.

**The diagnostic use of the chain is to find the weakest link.** For an ethical pattern, the chain survives being shown to the user at every link. For a manipulative one, there is exactly one link that **breaks on disclosure**, a point where, if the user could see it, the chain would snap.

| PATTERN | THE LINK THAT BREAKS ON DISCLOSURE |
|---|---|
| F-04 Drip Pricing | *Interface*, showing the total price up-front destroys it |
| G-01 Consent Asymmetry | *Interface*, equal-weight buttons destroy it |
| C-04 Streak Insurance | *Emotion*, "the anxiety you feel was manufactured by us" destroys it |
| F-14 Manufactured Urgency | *Trigger*, "this timer resets when you reload" destroys it |
| I-03 Frictionless Exit | **none**, full disclosure *strengthens* it |

This yields a test sharper than any severity rating, and we name it here:

**The Disclosure Test.** *Narrate the pattern's full DNA chain to the user in plain language. If the pattern still works, it is ethical or persuasive. If it collapses, it is manipulative.*

Manipulation is precisely the class of design that **cannot survive being explained to its subject.** That is not a moral claim. It is a structural one, and it is testable.

---

## 2. The Behavioral Genome

A product's signature across five axes, scored 0 to 10. Used at the head of every teardown.

| AXIS | QUESTION |
|---|---|
| **Attention** | How aggressively does it compete for focus? |
| **Habit** | How hard does it work to become automatic? |
| **Identity** | How much of the user's self-concept does it hold? |
| **Trust** | Would the user feel betrayed on learning how it works? *(high = trustworthy)* |
| **Autonomy** | How freely can the user leave, limit, or refuse? *(high = free)* |

**Read the last two axes carefully.** Attention, Habit, and Identity measure *grip*. Trust and Autonomy measure *whether the grip is legitimate*. A product may score high on all five: that is a product that has earned its hold. The pathological signature is **high grip with low Trust and low Autonomy**, a product that holds the user tightly and could not explain to them how.

Scores must be **justified per axis in prose.** A genome without justification is decoration.

**Warning against misuse.** The genome is a communication device, not a measurement. Nothing may be inferred from arithmetic on it: no averaging, no ranking of products by total. Two products with identical genomes may differ entirely. It exists to make a shape visible at a glance, and it is honest only if it is never asked to do more than that.

---

### 3. Behavioral Gravity

Friction is not a quantity. It is a **field with a direction.**

Every screen exerts force on the user along three vectors:

- **Attraction**, pulls the user toward an action
- **Friction**, resists an action
- **Momentum**, carries the user through an action they did not re-decide

The insight the existing literature misses: **friction is not bad and attraction is not good.** A confirmation dialog before deleting a year of work is friction that serves the user. Autoplay is momentum that does not. What matters is not the magnitude of the force but **whose goal it is aligned with.**

This gives the single most diagnostic measurement available on any consumer product, and it is cheap to take:

**The Asymmetry Ratio** = *(steps to start) ÷ (steps to stop)*

| RATIO | READING |
|---|---|
| ≈ 1 | Symmetric. The product is indifferent to your direction. **I-03.** |
| 2 to 3 | Mild retention design. Persuasive. |
| > 3 | The product is **structurally opposed to your leaving. F-09, H-01.** |

Signup: 2 clicks. Cancellation: 9 screens, a phone call, business hours only. The ratio is not a metaphor for intent. It **is** the intent, expressed in the only medium a product has.

Attraction and friction are both legitimate tools. The asymmetry between them, measured on the enter and exit paths of the same product, is the fingerprint of who the design is for.

---

### 4. Behavioral Debt

**The central longitudinal claim of this document, and the one most worth defending.**

Deceptive patterns work. That is why they exist. The Luguri and Strahilevitz experiments are decisive on this point: aggressive deceptive patterns raised acceptance of a dubious service from 11.3% to 41.9%, a 371% increase, and they found that under those conditions **price became immaterial**, decision architecture, not price, drove purchasing.

So the honest question is not *do they work?* They do. The question is: **what does using them cost the company that uses them?**

We propose the answer is a liability that behaves exactly like technical debt.

**Behavioral Debt** *is the accumulated, deferred cost of behavior extracted from users faster than value was delivered to them.*

#### The mechanics

Like technical debt, it is **borrowed against the future to pay for the present**. A deceptive pattern converts a unit of user trust into a unit of this quarter's metric. The metric is booked immediately. The trust is not booked at all, because no accounting system has a line for it.

Like technical debt, it **compounds**, and it compounds along a predictable path:

**The trap is at G.** The system's own instrumentation misreads the symptom as the cure. Falling engagement looks, on a dashboard, exactly like a need for more engagement mechanics. The company responds to behavioral debt by borrowing more.

This is the mechanism by which good companies with honest people ship terrible products. Nobody decided to be malicious. Every individual decision was a locally correct response to a metric. The metric was measuring the interest payment and reporting it as income.

#### Why it is invisible

Behavioral debt is not on the balance sheet, and the loop above explains why it never shows up as itself. Its symptoms are all attributed elsewhere:

| SYMPTOM | WHAT IT GETS CALLED INSTEAD |
|---|---|
| Users ignore notifications | "Notification fatigue." Fix: send more. |
| Users churn without complaint | "Poor product-market fit." |
| Trust scores fall | "Brand problem." Hand it to marketing. |
| Regulators arrive | "Compliance cost." Hand it to legal. |
| Users say they hate the app but keep using it | Read as engagement. **Booked as a win.** |

That final row is the most damning, and it is the clearest evidence the debt is real: **a product can be simultaneously at its most engaging and at its most hated.** Every existing metric reads that state as success. It is in fact the moment of maximum leverage, and maximum liability.

#### The prediction, and how to falsify it

Behavioral Debt is falsifiable, and this document should say how, or it is not a scientific claim.

**Prediction:** *hold engagement constant, and products with higher* **Opacity** *(Behavioral Cost Index) will show lower long-run retention, higher uninstall-after-lapse rates, and worse unprompted sentiment than products with lower Opacity.*

If products with high Opacity retain *just as well* over multi-year horizons, **the theory is wrong**, and the uncomfortable conclusion follows: deceptive patterns are simply free, and the only argument against them is a moral one.

We do not have the longitudinal data to settle this. Saying so is not a weakness of the document. Pretending otherwise would be.

#### Repayment

Debt can be paid down, and Family I is the instrument. Each entry in that family is a payment on a specific loan:

| DEBT INCURRED BY | REPAID BY |
|---|---|
| G-01, G-02, G-07 (consent extracted) | I-01 Honest Default, I-02 Symmetric Consent |
| F-09, H-01 (exit obstructed) | I-03 Frictionless Exit |
| A-01, A-02, A-13 (attention taken) | I-05 Sufficiency Nudge, I-06 Stopping Cue |
| F-11, H-09 (silent billing) | I-10 Unused-Value Alert |
| D-12, H-06 (user's work held hostage) | I-11 Portability by Design |

Note the shape of every repayment: **the company voluntarily surrenders a metric it could have kept.** That is what makes it repayment rather than marketing. Any "ethical design" initiative that costs the company nothing is not repaying anything.

---

RELATED

Behavioral Influence Model · Behavioral Cost Index (supplies the Opacity term on which Behavioral Debt depends) · TAXONOMY.md Family I

# Nine pattern entries

Nine of the 116, written to full template. One from each family, chosen so the valence thesis is tested at both poles: **I-03** costs the user nothing on any axis, **H-01** costs them on five.

A-01

## Infinite Scroll

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| A, Attention and Capture | Habit | **Persuasive → Manipulative** | ●●●●○ 4/5 |

### DEFINITION

Infinite scroll is a content-loading strategy in which the next batch of items is fetched and appended automatically as the user approaches the end of the current batch, so that the list never terminates. Its behavioral significance is not what it adds but what it **removes**: the page boundary. A paginated list ends, and its ending forces a decision ("do I click next?"). An infinitely scrolling list never ends, so continuing requires no decision at all, and stopping becomes the only act that requires one. The pattern relocates the burden of choice from continuation to cessation. Identifying it in the wild is trivial: if you can reach the bottom of the content and the bottom moves, it is infinite scroll.

### PATTERN DNA

**Goal** maximise session length and impressions per session → **Trigger** the scroll gesture itself, which is self-triggering and requires no external prompt → **Interface** a continuous list with pre-fetched content below the fold and no terminal boundary, often paired with a zero-latency loader (A-08) so no pause is visible → **Bias** the absence of a decision point; default-continuation, and (where content is algorithmically ordered) intermittent reinforcement → **Emotion** low-arousal absorption, then, characteristically, regret on exit → **Decision** none; the user does not decide to continue, they simply fail to decide to stop → **Behavior** extended, poorly time-tracked consumption → **Reward** the next item, whose value is variable and therefore reinforcing (C-02) → **Business metric** time in app, impressions, ad inventory served.

**The link that breaks on disclosure:** *Interface*. The chain does not depend on the user misunderstanding anything (most users know feeds are endless), which is why infinite scroll is not automatically manipulative. It depends on the *absence of a stopping cue*, and restoring the cue (I-06)

destroys the effect while leaving the honest part of the product intact. This is a rare and instructive case: the pattern's power lies in an omission rather than a deception.

### PSYCHOLOGICAL BASIS

The mechanism is **decision architecture, not persuasion**. Nothing is claimed to the user; a choice is simply never presented. Thaler and Sunstein's concept of choice architecture is the correct frame, and their later term **sludge** (friction that keeps people from what they *do* want) is the mirror image of what happens here: infinite scroll removes friction from a behavior the user does not, on reflection, want more of. Cite the framework, not an effect size. There is no defensible pooled "nudge effect size," and none is asserted here.

Two further mechanisms are commonly attached, with unequal support:

1. **Variable reward.** Where the feed is algorithmically ordered, each successive item is a draw from an unpredictable distribution, which is structurally a variable-ratio schedule (**Ferster & Skinner 1957**). The application of operant scheduling to feed consumption is an **analogy**, not a finding in that literature, and is labelled as such.
2. **Loss of temporal awareness.** The claim that infinite scroll degrades the user's estimate of elapsed time is plausible and is the basis of the adjacent pattern A-13, but no empirical source for it has been retrieved. CITATION NEEDED

The distinctive point, and the one the literature under-states: infinite scroll requires **no cognitive bias at all** to work. A perfectly rational agent, fully informed, still continues, because continuation is free and stopping is not. That makes it unusually resistant to the standard remedy of disclosure, and it is why this pattern needs a *structural* fix (I-06) rather than an informational one.

### INTERFACE EXPRESSION

- A vertically scrolling list with no terminal element and no page count.
- Pre-fetching, so the loading spinner is either absent or sub-perceptual: the boundary is not merely crossed, it is made invisible (A-08 Zero-Latency Feed).
- Absence of session-length cues: no clock, no "you have been here 40 minutes," no item counter (A-13 Temporal Disorientation).
- Frequently combined with autoplay (A-02) in video feeds, at which point even the scroll gesture becomes optional and consumption is fully passive.

### REAL-WORLD EXAMPLES

**Attribution.** Aza Raskin is widely credited with inventing infinite scroll and has publicly said he regrets it. This is now on the record in a proceeding rather than only in interviews: Raskin testified at the State of New Mexico's trial against Meta, where he was reported as saying he invented the feature and regrets a creation that traps users in endless scrolling and leads to wasted hours. This is reported testimony, and it is his own characterisation of his invention, not an independent finding of harm. The year of invention (commonly given as 2006) is **not verified against a retrieved primary source**. CITATION NEEDED

**Meta (Facebook, Instagram).** Infinite scroll is named explicitly in the bipartisan state attorneys-general litigation filed against Meta on 24 October 2023, which alleges that Meta designed and deployed

features encouraging addictive behavior in minors, specifically identifying algorithmic ordering, high-volume alerts and notifications, and infinite scroll through platform feeds. These are **allegations in pleadings**. They establish that the pattern is contested at law and that regulators regard the feed's endlessness as a designed property rather than an accident. They do not by themselves establish harm.

**Regulatory naming.** The European Parliament resolution of 12 December 2023 (adopted 545 in favour, 12 against, 61 abstentions) calls for assessment and possible bans on addictive techniques not covered by the Unfair Commercial Practices Directive, and names **"infinite scroll, default auto play, constant push and read receipt notifications"** among them. This is the clearest instance in the taxonomy of a regulator naming a specific interface mechanism rather than a category of deception.

No screenshot-level description of any current product interface is offered here, because interfaces change and memory of a past interface is not evidence about the present one (PROJECT.md §7 evidence rule).

#### BUSINESS RATIONALE

Session length is the numerator of nearly every attention-economy metric. Pagination imposes a decision point per page, and every decision point is an exit opportunity: a user who has to click "next" will sometimes decide not to. Infinite scroll removes those exits without removing any content. It costs almost nothing to implement, requires no additional content investment, and its benefit compounds with ad load, since more scroll depth is more inventory. Critically, it moves the metric **without any claim being made to the user**, which means it carries no obvious legal exposure under deception-based statutes. That is precisely why it is ubiquitous and why regulators have had to reach for new instruments (an addictive-design regime) rather than existing deception law.

#### BEHAVIORAL COST INDEX

Scored per frameworks/02. The scale measures the gap between cost paid and cost understood.

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 4 | The user spends materially more time than they intended, and would revise the behavior if the true elapsed total were made salient. The cost is real, large, and unaccounted. |
| Attention | 4 | Attention is *taken* rather than given: the continuation of the session is never re-decided, so no fresh grant of attention is ever made. |
| Memory | 0 | Nothing must be held in mind. This is the axis on which infinite scroll is genuinely cheap. |
| Trust | 2 | Users are broadly aware feeds are endless. On learning the mechanism they feel resigned rather than betrayed, which is a real and telling difference from, say, G-01. |
| Money | 0 | No direct charge. |
| Privacy | 0 | None intrinsic to the pattern. (The algorithmic ranking that usually accompanies it is A-09 and carries its own privacy cost.) |
| Emotion | 3 | Post-session regret is characteristic and is not disclosed. The user is not fully aware, in the moment, that the affect they will end with is dysphoric. |

**Max = 4 → Manipulative**, but only just, and the honest reading is that this pattern sits **on the boundary**. The two 4s are on Time and Attention, and both rest on the same claim: that awareness of the true cost would change behavior. That claim is arguable and is argued below rather than assumed. If a reader judges that users fully understand the time cost and accept it, the scores drop to 3 and the pattern is **Persuasive**. The instrument makes the disagreement precise instead of hiding it inside a severity number.

### VALENCE ANALYSIS

**The case for Persuasive.** Infinite scroll deceives nobody. It makes no claim, hides no price, and misstates no fact. Users know that feeds do not end; the endlessness is not a secret. It also serves a genuine user preference: pagination in a browsing context is annoying, and the pattern was invented as a usability improvement, not as a trap. A user who wants to browse a long list of search results, or a photo archive, is *better served* by continuous loading. In those contexts, the pattern is not merely persuasive but ethical, and the taxonomy must be able to say so.

**The case for Manipulative, which is the stronger one in feed contexts.** The valence turns on a single question: *does the design work because the user does not understand what is happening to them?* The answer is subtle. The user understands the mechanism perfectly. What the user does not have is a **stopping cue**, and the absence of that cue is not an oversight, it is the feature. The design does not exploit a false belief; it exploits the structural fact that a decision never presented is a decision never made. Under the Disclosure Test, narrating the chain does not restore the missing boundary. You can

tell the user exactly how infinite scroll works, and they will keep scrolling, because knowledge is not the thing they are short of.

This is why the pattern deserves its own treatment rather than being folded into deception. It defines a class of manipulation that **survives full disclosure**, and that class is invisible to every taxonomy organised around the question "how is the user being deceived?" (Brignull; Mathur et al. 2019; Gray et al. 2018). The correct diagnostic here is not honesty but **whose goal the momentum serves** (Behavioral Gravity, frameworks/03).

**Where the line actually falls.** Infinite scroll is *Persuasive* when the content is finite and user-requested (a search result set, a user's own photo library, a chronological feed of people they chose to follow) and *Manipulative* when it is combined with algorithmic ordering optimised for engagement (A-09), because the compound has no terminus **in principle**: the supply of content is unbounded and is selected to defeat cessation. The mechanism is the same; the compound is not.

#### ETHICAL ALTERNATIVE

The legitimate goal is a browsing experience without the annoyance of pagination. That goal does not require the removal of every boundary.

1. **I-06 Stopping Cue.** Restore an end. A designed terminus ("you are all caught up") gives the user a natural exit at zero cost to the honest part of the session. This is not hypothetical: end-of-feed markers have shipped in mainstream feed products. It forfeits scroll depth, which is exactly what makes it a genuine concession rather than a marketing gesture.
2. **Bounded batches with an explicit continue.** Load 50 items, then require one deliberate action to load 50 more. This preserves the anti-pagination benefit (no page reloads, no lost place) while re-introducing a decision point. The cost to the user is one tap. The cost to the business is a measurable exit rate, which is the point.
3. **I-04 Usage Transparency.** Surface elapsed session time and items consumed inside the feed, not buried in a settings screen. A product confident that its feed is worth the time can afford to show the time.
4. **I-05 Sufficiency Nudge.** After a threshold the user set, actively suggest they stop. This is the only remedy that addresses the structural problem rather than the informational one, because it supplies the missing decision point on the user's behalf.
5. **I-12 Consumption Budget.** A user-set limit that the product enforces *against its own engagement interest*.

The test of good faith across all five: does the intervention cost the company a metric it could have kept? If not, it is not a remedy.

#### REGULATORY STATUS

- **No jurisdiction currently prohibits infinite scroll.**
- **European Union.** The European Parliament's own-initiative resolution of 12 December 2023 on addictive design of online services explicitly names infinite scroll and asks the Commission to assess and consider banning it where it is not already covered by the Unfair Commercial Practices Directive.

A Parliament resolution is a political request, not a binding rule. Separately, "dark patterns" were codified into EU law in 2022 through the Digital Services Act, the Digital Markets Act, and the Data Act proposal (Gray et al. 2024). The DSA is understood to prohibit deceptive interface design by online platforms, but the specific provision and its wording were **not retrieved for this document** and are therefore CITATION NEEDED. The substantive point stands regardless: infinite scroll is not deception, so a deception-based prohibition would not obviously reach it, and that is precisely the gap the resolution identifies.

- **United States.** No federal rule. Infinite scroll appears as alleged conduct in the multistate attorneys-general litigation against Meta (filed 24 October 2023 in federal court by a bipartisan coalition of states, with further suits filed in state courts), brought under state consumer-protection statutes and COPPA. Unadjudicated.
- **Note on the regulatory difficulty.** Infinite scroll is the clearest example of why deception-based law struggles with behavioral design. Nothing false is said. Any effective regulation must therefore reach *architecture* rather than *representation*, which is a materially harder thing for a consumer-protection statute to do.

### RELATED PATTERNS

- **A-02** Autoplay, the same removal of a decision point applied to video.
- **A-08** Zero-Latency Feed, which conceals the batch boundary that would otherwise act as a residual cue.
- **A-09** Algorithmic Curation, the compound that converts this pattern from Persuasive to Manipulative.
- **A-13** Temporal Disorientation, the removal of the clock that would let the user price the Time cost.
- **C-02** Variable Reward, the reinforcement layer beneath an algorithmically ordered feed.
- **I-05** Sufficiency Nudge, **I-06** Stopping Cue, **I-04** Usage Transparency, **I-12** Consumption Budget: the ethical counterparts.

### SOURCES

Verified project sources (see research/sources/):


- Thaler, R.H., & Sunstein, C.R. (2021). *Nudge: The Final Edition*. New Haven: Yale University Press. (Choice architecture as a framework; "sludge." Cited as a framework, **not** as an effect size: the pooled nudge effect is disputed under publication-bias correction.)
- Ferster, C.B., & Skinner, B.F. (1957). *Schedules of Reinforcement*. New York: Appleton-Century-Crofts. (Book-level citation; the extension to feed consumption is labelled an analogy.)
- Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A.L. (2018). "The Dark (Patterns) Side of UX Design." CHI '18, Paper 534. DOI 10.1145/3173574.3174108.
- Mathur, A., et al. (2019). "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites." *PACM HCI* 3(CSCW), Art. 81. DOI 10.1145/3359183. (Cited for what its taxonomy does *not* cover: architectural, non-deceptive capture.)
- Gray, C.M., Santos, C.T., Bielova, N., & Mildner, T. (2024). "An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building." CHI '24. DOI 10.1145/3613904.3642436. (Source for the DSA/DMA/CPRA codification context.)

Sources:


- European Parliament, "New EU rules needed to address digital addiction," press release 20231208IPR15767. https://www.europarl.europa.eu/news/en/press-room/20231208IPR15767/new-eu-rules-needed-to-address-digital-addiction (vote 545/12/61; "infinite scroll, default auto play, constant push and read receipt notifications" quoted verbatim from the release).
- KOB 4 (New Mexico), "Inventor of infinite scroll says he regrets it at New Mexico Meta trial." https://www.kob.com/new-mexico/inventor-of-infinite-scroll-says-he-regrets-it-at-new-mexico-meta-trial/ (Raskin's testimony; publication date not stated on the retrieved page).
- CNBC, "Meta sued by 42 attorneys general alleging Facebook, Instagram features are addictive and target kids," 24 October 2023. https://www.cnbc.com/2023/10/24/bipartisan-group-of-ags-sue-meta-for-addictive-features.html (infinite scroll named among the alleged features).
- Office of the New York Attorney General, press release on the multistate Meta suit. https://ag.ny.gov/press-release/2023/attorney-general-james-and-multistate-coalition-sue-meta-harming-youth


**Attribution status.** TAXONOMY.md records A-01's attribution to Aza Raskin as **pending**. The New Mexico trial reporting above corroborates that Raskin himself claims the invention, but a primary source for the invention itself (a 2006 artifact, patent, or contemporaneous publication) has **not** been retrieved. The attribution therefore remains ATTR PENDING and the invention year is CITATION NEEDED.

B-09

# Permission Priming

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| B, Onboarding and Activation | Onboarding | **Persuasive → Manipulative** | ●●●○○ 3/5 |

## DEFINITION

Permission priming is the insertion of an application-controlled screen immediately before an operating-system permission dialog, for the purpose of raising the probability that the user grants the permission. The OS dialog (for notifications, location, contacts, camera, tracking) is typically a **one-shot** resource: on the major mobile platforms it can be presented only once per permission, and a denial is expensive to reverse because the user must then be sent into system settings. The pre-prompt exists to protect that one shot. It is not itself a permission request, it is a *screening* request, and its distinguishing feature is that a "no" costs the app nothing: the user who declines the soft ask is simply never shown the hard one, and can be asked again later. Identifying it in the wild: any in-app screen that asks whether you would *like* to enable something, styled by the app rather than the OS, and immediately followed by the real system alert if you say yes.

## PATTERN DNA

**Goal** maximise the opt-in rate for a permission that a growth or monetisation loop depends on → **Trigger** a chosen moment in onboarding, usually immediately after a first success, when goodwill is highest → **Interface** an app-styled modal, often with an illustration and a benefit statement, offering a positive option and a soft deferral ("Not now") rather than a hard refusal → **Bias** consistency and commitment: a small prior agreement makes the subsequent larger one harder to refuse (B-10); framing, since the benefit is stated and the cost is not → **Emotion** anticipated benefit, mild social compliance → **Decision** "yes, that sounds useful" → **Behavior** the user taps through the OS dialog, which now arrives pre-endorsed → **Reward** the promised feature, sometimes → **Business metric** notification opt-in rate (which drives A-04 and retention), contacts opt-in rate (which drives E-03 and viral growth), or tracking opt-in rate (which drives ad revenue).

**The link that breaks on disclosure:** *Interface*, in the manipulative form only. "We are showing you this screen first so that if you say no, we can preserve our one chance and ask you again in a week" is a sentence that, if printed on the modal, would sharply reduce compliance. In the persuasive form there is no such sentence to hide: the pre-prompt is genuinely explaining *why* the permission is needed, which is information the OS dialog is too small to carry.

## PSYCHOLOGICAL BASIS

1. **Commitment and consistency.** The soft ask converts the hard ask into a confirmation of a decision the user has already made. This is Cialdini's consistency principle (one of the **seven**, not six, in the 2021 expanded edition). Replication across Cialdini's principles is mixed and principle-dependent; consistency is not among the strongest-evidenced and should not be presented as settled. CITATION NEEDED for a replication-grade effect size on commitment-and-consistency specifically.

2. **Kairos and the timing of the ask.** Fogg (2009) proposes that behavior is a product of three factors, motivation, ability, and triggers, and introduces *kairos*, the opportune moment to persuade. Permission priming is a kairos device: it does not change the permission, it changes *when* the permission is asked, moving it to a point of peak motivation (just after a first success) rather than a point of peak suspicion (app launch). Note carefully: Fogg's 2009 paper contains **no equation**. The formulation "B = MAT" is a community retrofit and does not appear in the paper. Fogg renamed "trigger" to "prompt" in late 2017.
3. **Framing.** The pre-prompt states a benefit and omits a cost. Prospect theory's insight that valuation is defined over gains and losses relative to a reference point (Kahneman & Tversky 1979, replicated by Ruggeri et al. 2020) predicts that "get reminders so you never miss a lesson" and "let us interrupt you at our discretion" will be valued very differently despite describing the same grant.

The specific claim that pre-prompts raise opt-in rates, which is the entire practitioner rationale for the pattern, is an **industry claim** with no retrieved academic support. CITATION NEEDED for any opt-in lift figure. Practitioner blogs quote such figures freely; none were traceable to a primary source in this research and none are printed here.

#### INTERFACE EXPRESSION

- An app-styled modal, visually distinct from the OS alert, appearing in onboarding.
- Two options with **asymmetric affordances**: a prominent affirmative ("Enable notifications") and a low-contrast deferral ("Maybe later"). Note that the deferral is usually *not* a refusal: it preserves the app's right to ask again.
- Benefit-framed copy naming a user-facing gain, not the business function the permission actually serves.
- Placement immediately after a first completed action, so the request lands in a moment of gratitude.
- In the manipulative form, the modal is styled to **resemble the system dialog**, blurring which entity is asking. This is the single clearest tell.
- In the most aggressive form, the pre-prompt is not dismissible, or is re-presented on every launch until the user yields. At that point the pattern has become G-05 Permission Nagging.

#### REAL-WORLD EXAMPLES

**The platform rules are the best-documented artifact here, and they exist because the pattern exists.** Apple's App Store Review Guidelines, retrieved directly, contain two provisions that bound this pattern:

- Guideline **5.1.1(iv) Access**, verbatim: "*Apps must respect the user's permission settings and not attempt to manipulate, trick, or force people to consent to unnecessary data access. For example, apps that include the ability to post photos to a social network must not also require microphone access before allowing the user to upload photos. Where possible, provide alternative solutions for users who don't grant consent.*"
- Guideline **5.1.2(i) Data Use and Sharing**, verbatim: "*You must receive explicit permission from users via the App Tracking Transparency APIs to track their activity. ... Your app may not require users to*

> *enable system functionalities (e.g. push notifications, location services, tracking) in order to access functionality, content, use the app, or receive monetary or other compensation, including but not limited to gift cards and codes.*"

Two things follow. First, the existence of an explicit platform prohibition on **conditioning access or compensation on a permission grant** is direct evidence that apps were doing exactly that. Second, the rules constrain *coercion* and *incentivisation* but do **not** prohibit a benefit-framed pre-prompt, which is why the pattern remains near-universal and is generally considered legitimate practice by the platform itself.

**App Tracking Transparency.** Apple's ATT prompt is the highest-stakes one-shot dialog in consumer software, because a denial removes the identifier that mobile advertising attribution depends on. The pre-prompt in front of it is therefore the most economically consequential instance of this pattern. Apple's own guidance, per 5.1.2(i), is that consent must be obtained through the ATT API and may not be purchased or coerced.

No specific app's onboarding screen is described here. Onboarding flows are A/B tested continuously and change without notice; describing a remembered screen would violate the evidence rule in PROJECT.md §7.

#### BUSINESS RATIONALE

Permissions are the gates on the highest-leverage growth loops in mobile software. Notification permission gates re-engagement (A-04, C-06). Contacts permission gates viral growth (E-03, E-07). Tracking permission gates ad monetisation and attribution. Each is a **one-shot** ask whose denial is close to permanent, because the recovery path runs through the OS settings app, where conversion is near zero.

Given that structure, a rational company will do exactly what this pattern does: interpose a cheap, reversible screening step in front of an expensive, irreversible one. The pre-prompt costs a modal. A denied system dialog costs the loop. Note that the pattern is a *rational response to a platform constraint*, and the constraint was itself introduced to protect users. This is a recurring dynamic worth naming: **a protection that is one-shot invites a pattern that optimises the one shot**, and the protection thereby creates the pattern it was meant to prevent.

#### BEHAVIORAL COST INDEX

Scored per frameworks/02. Two implementations, because the range matters.

| AXIS | HONEST PRE-PROMPT (STATES THE REAL REASON, "NO" IS A REAL NO) | MANIPULATIVE PRE-PROMPT (MIMICS SYSTEM UI, BENEFIT-ONLY FRAMING, RE-ASKS) |
|---|---|---|
| Time | 1 | 2 |
| Attention | 1 | 2 |
| Memory | 0 | 1 |
| Trust | 1 | 3 |
| Money | 0 | 0 |
| Privacy | 2 | 4 |
| Emotion | 0 | 2 |

**Honest column.** *Time 1 / Attention 1:* one extra tap, which the user would accept if asked plainly, because the screen carries genuinely useful context the OS dialog cannot. *Memory 0:* nothing to retain. *Trust 1:* being asked politely before being asked formally is, if anything, a courtesy. *Privacy 2:* a real disclosure occurs, the user is aware of it, and the friction is mild. **Max = 2 → Ethical**, and this is a case where the taxonomy must be willing to say that a growth-motivated pattern is not a harm.

**Manipulative column.** *Trust 3:* the user is not fully aware that the modal is the app, not the system, nor that "Maybe later" is an option the app engineered so it may ask again. *Privacy 4:* the permission granted is broader than the benefit stated, the user is unaware of the gap, and awareness would change the decision. This is the load-bearing score. *Emotion 2:* mild pressure. **Max = 4 → Manipulative.**

**Two discrepancies with TAXONOMY.md, recorded rather than smoothed over.** First, the taxonomy gives B-09 a valence *range* of Persuasive → Manipulative, but the honest column above scores a maximum of 2, which under the BCI's derivation rule (frameworks/02: Ethical = every axis ≤ 2) is **Ethical**. The instrument therefore says the declared floor of the range is one band too high. Second, the pattern's severity ceiling is 3 in TAXONOMY.md. This scoring suggests the ceiling is **too low** and should be 4, since a pre-prompt that mimics system UI in service of a tracking grant produces exactly the "unaware, and awareness would change behavior" condition that defines a 4. Recorded here as a discrepancy to resolve when the ceilings are re-derived from the BCI, which TAXONOMY.md open question 4 already flags as necessary.

### VALENCE ANALYSIS

**The case for Ethical, which is real and must be conceded.** The OS permission dialog is a terrible piece of communication. It is short, it is generic, it is written by the platform and not by the app, and it arrives with no context. A user asked "Allow Notifications?" with no explanation is not making an informed decision, they are making a defensive one. A pre-prompt that says "*we will send you one reminder per day at a time you choose, and nothing else*" gives the user strictly more information than the OS gives them, and a better-informed decision is not manipulation, in either direction. If the pre-prompt honestly describes what the permission will be used for, and if declining it is a genuine decline rather than a deferral, then this pattern **improves** consent quality. That is the strongest defence available for any pattern in Family B, and it should not be waved away.

**The case for Persuasive.** The company chooses the moment. Asking at peak goodwill rather than at a neutral moment is not deception, but it is not neutral either: it is a deliberate exploitation of *kairos*. The

user's answer is a function of when they were asked, and only the company knows that. This is steering with the user's eyes open, which is the definition of the persuasive band.

**The case for Manipulative, and where the line actually is.** Three tells, any one of which is sufficient:

1. **The modal impersonates the system.** If the user cannot tell whether they are talking to the app or to the OS, the informational benefit of the pre-prompt inverts into a harm: the user believes they are answering a platform question when they are answering an interested party's question.
2. **The decline is not a decline.** A "Maybe later" that exists solely so the app can re-ask is a pre-prompt whose function is not to inform but to *ration the one-shot dialog*. The user believes they have refused. They have been deferred. This is the exact structure that becomes G-05 Permission Nagging on repetition.
3. **The stated benefit is not the actual use.** "Enable notifications so you never miss a message from a friend" followed by promotional pushes is a misrepresentation, and one that existing deception law already reaches.

Under the Disclosure Test, the honest pre-prompt is unusually robust: telling the user everything about it makes it *work better*, because its whole function is to inform. The manipulative pre-prompt collapses on the first sentence of disclosure. That asymmetry, within one pattern, is the cleanest small demonstration of the valence thesis outside C-01.

#### ETHICAL ALTERNATIVE

The legitimate goal, a well-informed permission decision made at a moment when the user can understand the request, is achievable without any of the harms.

1. **Ask in context, not in onboarding.** Request notification permission at the moment the user does something that *creates* a reason for a notification (sets a reminder, follows a thread). The request then explains itself, and the opt-in is high for the right reason.
2. **Make the pre-prompt do real work.** State the specific use, the frequency, and what will *not* be sent. A pre-prompt that carries a promise the app can be held to is an asset, not a trick.
3. **Make "no" a real no.** If the user declines the soft ask, do not re-present it. Instead, place a persistent, findable control in the product ("Turn on reminders") that they can use when they choose. This forfeits opt-in rate, which is precisely what makes it a genuine concession.
4. **Never mimic the system chrome.** Style the pre-prompt unmistakably as the app. Ambiguity here is never accidental.
5. **Symmetric affordances** (I-02): the affirmative and the decline get equal visual weight. A pre-prompt with a grey "Not now" and a saturated "Allow" has already conceded that it is not trying to inform.

#### REGULATORY STATUS

- **Platform rules, not law, are the binding constraint here, and they are documented.** Apple's App Store Review Guidelines 5.1.1(iv) prohibit manipulating, tricking, or forcing consent to unnecessary data access, and 5.1.2(i) prohibits conditioning functionality, content, or compensation on enabling push notifications, location, or tracking. Violation is an app-review rejection, which is a more immediate sanction than most statutes provide.

- **European Union.** Where the permission covers personal data, GDPR requires consent to be freely given, specific, informed and unambiguous. A pre-prompt whose stated purpose differs from the actual processing purpose is a purpose-limitation problem, not merely a design one. "Dark patterns" were codified into EU law in 2022 through the Digital Services Act, the Digital Markets Act, and the Data Act proposal (Gray et al. 2024). CITATION NEEDED for the specific GDPR article text and for the specific DSA provision on deceptive interface design; **neither was retrieved in this research**, and neither is paraphrased here.
- **United States.** No pattern-specific federal rule. A pre-prompt whose stated benefit misdescribes the actual use is within reach of FTC Act §5 unfair-and-deceptive-practices authority; Luguri & Strahilevitz (2021) argue that many deceptive patterns already violate federal and state UDAP statutes without new legislation.

### RELATED PATTERNS

- **B-10** Commitment Escalation, the mechanism the pre-prompt runs on.
- **B-11** Contact Ingestion Prompt, the highest-stakes application of it, and **E-03** Contact Harvesting, the loop it feeds.
- **G-05** Permission Nagging, what this pattern becomes on repetition. The boundary between B-09 and G-05 is exactly one re-ask.
- **G-08** Consent Nudge, the visual-weight asymmetry that usually accompanies it.
- **G-04** Bundled Consent, **G-12** Ambiguous Wording, adjacent consent-quality failures.
- **A-04** Notification Cascade, the downstream use of the permission this pattern secures.
- **I-02** Symmetric Consent, the ethical counterpart.

### SOURCES

Verified project sources (see research/sources/):


- Fogg, B.J. (2009). "A Behavior Model for Persuasive Design." Persuasive '09, Article 40. DOI 10.1145/1541948.1541999. (Motivation, ability, triggers; *kairos*. **The paper contains no equation**; "B=MAT" is a later community retrofit and is not used here. "Trigger" was renamed "Prompt" by Fogg in late 2017.)
- Cialdini, R.B. (2021). *Influence, New and Expanded: The Psychology of Persuasion*. New York: Harper Business. (Consistency, one of **seven** principles. Per-principle replication is mixed; not presented as settled.)
- Kahneman, D., & Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2), 263-291. DOI 10.2307/1914185; with Ruggeri, K., et al. (2020), *Nature Human Behaviour* 4(6), 622-633, DOI 10.1038/s41562-020-0886-x (multinational replication).
- Luguri, J., & Strahilevitz, L.J. (2021). "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006.
- Gray, C.M., et al. (2018). "The Dark (Patterns) Side of UX Design." CHI '18. DOI 10.1145/3173574.3174108. (Nagging as a strategy category.)


Sources:

- Apple, *App Store Review Guidelines*, sections 5.1.1(iv) and 5.1.2(i). https://developer.apple.com/app-store/review/guidelines/ (both provisions quoted verbatim above from the retrieved text).
- Apple, *User Privacy and Data Use*. https://developer.apple.com/app-store/user-privacy-and-data-use/ (App Tracking Transparency requirements).


**Unverified and therefore omitted:** all practitioner-reported opt-in "lift" percentages for pre-permission prompts. None was traceable to a primary source.

C-01

# Streak

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| C, Habit and Reinforcement | Habit | **Ethical → Manipulative (implementation-dependent)** | ●●●●○ 4/5 |

## DEFINITION

A streak is a running count of consecutive periods (usually days) in which the user performed a qualifying action, displayed persistently and reset to zero on any lapse. Its defining property is that the counter has no intrinsic value: nothing the user owns, learned, or achieved is destroyed when the number returns to zero. The number is valuable to the user *only because it is fragile*, and its fragility is a design decision, not a fact about the world. A streak is therefore the purest available case of a mechanism whose ethical status cannot be read off the mechanism. The same counter can encode a commitment device the user chose, or a manufactured anxiety the product then sells the user a remedy for.

This entry is the pivot of the whole taxonomy. If the valence thesis is wrong anywhere, it is wrong here.

## PATTERN DNA

**Goal** raise D1/D7 retention and session frequency → **Trigger** a scheduled push notification at the user's historical usage window (C-06), or the visible counter itself → **Interface** a persistent numeric badge, often with a flame or equivalent icon, plus a countdown or "you still have N hours" warning → **Bias** loss aversion: the counter is coded as a possession, so a lapse is framed as a loss rather than as a non-gain → **Emotion** anticipatory anxiety, then relief on completion → **Decision** "do the minimum qualifying action now" → **Behavior** a short, often low-quality session performed to satisfy the counter → **Reward** the counter increments; in some products a celebratory animation fires → **Business metric** consecutive-day active users, D7/D30 retention, and (where C-04 Streak Insurance exists) direct revenue.

**The link that breaks on disclosure:** *Emotion*, but only in the manipulative implementation. Telling a self-directed language learner "we display this counter because it helps you keep a habit you told us you wanted" does not weaken the streak. Telling a teenager "the dread you feel about your Snapstreak is an artifact we built, and nothing real is lost" does weaken it. Same mechanism, opposite result under the Disclosure Test (Behavioral Pattern DNA, frameworks/03).

## PSYCHOLOGICAL BASIS

Three mechanisms are commonly invoked. Their evidential status is not equal, and the difference matters.

1. **Loss aversion.** The value function is "generally steeper for losses than for gains" (Kahneman & Tversky 1979). A streak works by moving the reference point: once the counter exists, the user is no longer choosing whether to gain a day, they are choosing whether to lose thirty. Prospect theory is

one of the sturdier items in this literature: Ruggeri et al. (2020) replicated 94% of items across 4,098 participants in 19 countries. This is the strongest available foundation for the streak.

2. **The goal gradient / completion pull.** The tendency to resume an interrupted task is the **Ovsiankina effect** (pooled resumption rate 67.0%, Ghibellini & Meier 2025), *not* the Zeigarnik effect, whose memory advantage does not replicate (pooled recall ratio 0.99). Product writing routinely cites Zeigarnik here and is wrong to. The honest claim is about resumption, not recall, and resumption is exactly what a streak solicits.
3. **Reinforcement scheduling.** A streak is a *fixed-ratio, continuous* reinforcement schedule (one action, one increment), which is the schedule most vulnerable to rapid extinction once reinforcement stops. This is why streak products bolt variable-reward layers (C-02) onto them. The canonical source is **Ferster & Skinner (1957)**, and the claim that operant schedules explain compulsive app use is an *analogy* drawn from that literature, not a finding within it. It is labelled as such here.

The specific claim that streaks raise retention is an industry claim, not an academic one. CITATION NEEDED for any effect size on streaks in particular.

#### INTERFACE EXPRESSION

A persistent counter, usually in the top bar or profile, rendered with an icon that implies heat, life, or fragility. Three sub-components recur:

- **A visible expiry.** The counter is paired with time pressure ("your streak ends in 4 hours"). Without expiry the counter is a record; with expiry it is a threat.
- **An escalating notification ladder.** Reminders intensify as midnight approaches. In the manipulative form the copy migrates from informational ("keep your streak alive") to relational or accusatory (C-14 Guilt Messaging).
- **A repair mechanism.** A freeze, a wildcard, or a paid restore. The moment this exists, the product is monetising an anxiety it created (C-04).

#### REAL-WORLD EXAMPLES

**Duolingo (language learning).** Duolingo's streak is the most publicly discussed implementation in consumer software. Its documented architecture includes the streak counter itself, a "Streak Freeze" that protects the count through a missed day, and streak repair. The specific retention figures that circulate for these features (churn reduction percentages, DAU multiples) come from third-party marketing blogs, not from Duolingo's own filings, and are **not** reproduced here as fact. CITATION NEEDED for any Duolingo streak effect size.

What *is* observable and uncontroversial: the streak is opt-out-able in the sense that a user who does not care about it can ignore it and still use the product; the qualifying action is the product's core value-delivering action (a lesson); and the user came to the product with a stated goal ("learn Spanish") that daily practice genuinely serves.

**Snapchat (messaging).** Snapchat's "Snapstreak" counts consecutive days on which two users have exchanged snaps. The qualifying action is not a value-delivering action; it is *any* snap, and the streak is held jointly, so a lapse harms another person (D-11 Group Obligation). Snapstreaks are named explicitly in state attorney-general litigation against Snap Inc.: the Arkansas Attorney General's suit alleges that

disappearing messages, Snapstreaks, and frequent notifications together created addictive feedback loops in adolescents, and the Texas Attorney General's suit likewise cites Snapstreaks among the incentives to daily use alleged to harm minors. These are **allegations in pleadings, not adjudicated findings**, and are cited here as evidence that the pattern is contested at law, not as proof of harm.

**Regulatory framing.** The European Parliament's December 2023 resolution on addictive design (adopted 545 to 12, with 61 abstentions) does not name streaks, but it names the family they belong to and calls for assessment and possible bans on techniques including infinite scroll, default autoplay, and constant push and read-receipt notifications.

### BUSINESS RATIONALE

A streak converts a probabilistic behavior (will the user come back?) into a scheduled one. It is the cheapest known device for raising consecutive-day active users, and consecutive-day usage is the single strongest leading indicator of long-run retention in most consumer subscription businesses CITATION NEEDED. It costs nothing to build, requires no content, and compounds: the longer the streak, the greater the switching cost, so the marginal retention value of the counter *increases* with tenure. When paired with a paid repair (C-04), it also converts anxiety directly into revenue, which is why the repair almost always ships.

A rational company ships this even with no intent to harm. That is the point of the Behavioral Debt argument (frameworks/03): the metric books the gain immediately and has no line for the cost.

### BEHAVIORAL COST INDEX

Scored twice, because the pattern is identical and only the implementation differs. Scores follow the BCI scale in frameworks/02 (a score measures the **gap between cost paid and cost understood**, not magnitude).

| AXIS | SELF-DIRECTED STREAK (LEARNING APP) | SOCIALLY-BOUND STREAK WITH GUILT COPY AND PAID REPAIR |
|---|---|---|
| Time | 1 | 2 |
| Attention | 1 | 3 |
| Memory | 1 | 1 |
| Trust | 0 | 3 |
| Money | 0 | 4 |
| Privacy | 0 | 0 |
| Emotion | 1 | 5 |

**Left column, justification.** *Time 1:* the daily session is short and the user would have accepted it if asked plainly, because it is the activity they signed up to do. *Attention 1:* one reminder, at a time the user effectively chose. *Memory 1:* the user must remember to practise, which is the point. *Trust 0:* nothing is concealed. *Money 0:* no charge. *Privacy 0:* no disclosure. *Emotion 1:* mild pressure, of a kind the user would endorse on reflection. **Max = 1 → Ethical.**

**Right column, justification.** *Time 2:* sessions become obligatory and are performed for the counter rather than for value, but the user is aware of this. *Attention 3:* the notification ladder claims vigilance

the user did not budget for. *Memory 1:* unchanged. *Trust 3:* the user is not fully aware that the deadline is arbitrary and company-set. *Money 4:* the user pays to repair a loss whose stakes the company invented, and would very plausibly not pay if that were made salient. *Emotion 5:* this is the crux. The product manufactures an anxiety with **no referent in reality** (nothing is genuinely lost by missing a day) and then sells the remedy for it. The design works *precisely because* the user has not noticed that the thing they fear losing was created by the company now charging them to protect it. **Max = 5 → Manipulative.**

Note that the instrument, not the analyst, produces the two verdicts. That is the claim the BCI is making.

#### VALENCE ANALYSIS

The honest version, and it has to concede more than a moralising version would.

**The case that the self-directed streak is ethical.** The user arrived with a goal that daily repetition genuinely serves. The counter is a *commitment device*: a self-imposed cost on future defection, of exactly the kind a rational agent with a self-control problem would choose to install. The business interest (daily active users) and the user interest (learning a language) are not merely compatible, they are the same variable. Crucially, the streak survives the Disclosure Test. You can tell the user precisely how and why it works, and it keeps working, because the user endorses the mechanism on reflection. That is the operational definition of persuasion rather than manipulation.

**The case against, which must be made.** Even the benign streak subtly *substitutes* the counter for the goal. A user optimising for the streak will take the shortest qualifying action, not the most educationally valuable one. The metric displaces the objective it was proxying. And the moment a repair is sold, the product acquires a financial interest in the user's anxiety, which is a conflict of interest regardless of how gently it is exercised. Duolingo is not a saint in this story; it is a product that has built freeze and repair mechanics, and those mechanics sit at the boundary of C-04. The correct verdict on the self-directed streak is **Ethical, trending Persuasive**, not Ethical unqualified.

**The case that the socially-bound streak is manipulative.** Three properties, jointly, cross the line. (1) The qualifying action carries no value: sending a blank snap to preserve a count is pure ritual, and the product knows it. (2) The obligation is *interpersonal*: lapsing harms a friend, so the cost of exit is paid in a currency the company does not own and did not create. (3) The anxiety is monetisable, and where it is monetised, the company is on both sides of the trade. Under the Disclosure Test the chain snaps at *Emotion*: a user who fully internalised "nothing is actually lost, and the dread is an artifact" would not pay to prevent the loss, and might not perform the ritual at all.

**What this establishes.** The two implementations share a mechanism, an interface, a bias, and a business metric. They differ at exactly two links of the DNA chain: whether the qualifying action delivers the value the user came for, and whether the emotion survives disclosure. A taxonomy that files "streak" permanently under dark patterns cannot state that difference. A taxonomy that files it under gamification cannot state it either. This is what the valence axis is for.

#### ETHICAL ALTERNATIVE

Not "remove the streak." The legitimate business goal (habit formation, which the user also wants) is real and has a non-harmful solution.

1. **Make the qualifying action the value-delivering action.** A streak should be impossible to satisfy without receiving the benefit the user came for. This single constraint kills the blank-snap ritual.
2. **Let the user set the terms.** A "three days a week" streak, chosen by the user, is a commitment device. A daily streak, imposed by the product, is a schedule the company set for its own metric. This is I-12 Consumption Budget applied to habit design.
3. **Give the repair away, or do not sell it.** A free freeze is a kindness. A paid freeze is a market in an anxiety the seller manufactured. If the repair must exist, it should be free and automatic, which forfeits revenue: per frameworks/03, that forfeiture is what makes it repayment rather than marketing.
4. **Cap the guilt.** Reminder copy should be informational and should de-escalate, not escalate, as the deadline nears. Escalation is C-14.
5. **Show the user what the streak actually bought them.** Replace "37 days" with "37 days, 19 lessons, 340 new words." This is I-04 Usage Transparency, and it is a direct test of good faith: a product whose streak is not proxying real value cannot make this substitution.

### REGULATORY STATUS

- **No statute directly regulates streaks anywhere, as of this writing.**
- **European Union.** The European Parliament's resolution of 12 December 2023 on addictive design of online services (adopted 545/12/61) asks the Commission to assess and consider banning addictive techniques not covered by the Unfair Commercial Practices Directive, explicitly naming infinite scroll, default autoplay, and constant push and read-receipt notifications. Streaks are not named. A resolution is not law; it is an own-initiative request to the Commission.
- **United States.** No federal rule. Streaks appear as *alleged conduct* in state consumer-protection litigation, notably the Arkansas and Texas Attorneys General suits against Snap Inc., which name Snapstreaks among features alleged to be addictive by design. These are unadjudicated allegations.
- **General law.** Where a streak is paired with a paid repair and the urgency is misrepresented, existing unfair-and-deceptive-practices law is available in principle: Luguri & Strahilevitz (2021) argue that many deceptive patterns already violate federal and state UDAP statutes without new legislation.

### RELATED PATTERNS

- **C-04** Streak Insurance, the monetisation of the anxiety this pattern creates. The single most diagnostic co-occurrence.
- **C-14** Guilt Messaging, the escalation path of the reminder ladder.
- **C-06** Trigger Scheduling, which supplies the notification timing.
- **C-12** Reward Escalation and **C-10** Tier Progression, adjacent loss-framing mechanisms.
- **D-11** Group Obligation, which converts a personal streak into an interpersonal one and is the specific ingredient that makes Snapstreak-type implementations manipulative.
- **C-05** Goal Gradient, **C-13** Completion Tension, the resumption mechanics underneath.
- **I-12** Consumption Budget and **I-04** Usage Transparency, the ethical counterparts.

### SOURCES

Verified project sources (see research/sources/):


- Kahneman, D., & Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2), 263-291. DOI 10.2307/1914185. (Value function steeper for losses than gains.)
- Ruggeri, K., et al. (2020). "Replicating patterns of prospect theory for decision under risk." *Nature Human Behaviour* 4(6), 622-633. DOI 10.1038/s41562-020-0886-x. (4,098 participants, 19 countries; 94% of items replicated.)
- Ghibellini, R., & Meier, B. (2025). "Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects." *Humanities and Social Sciences Communications* 12, art. 962. DOI 10.1057/s41599-025-05000-w. (Zeigarnik recall ratio 0.99; Ovsiankina resumption 67.0%.)
- Ferster, C.B., & Skinner, B.F. (1957). *Schedules of Reinforcement*. New York: Appleton-Century-Crofts. (Cited at book level; no figures attached, per the Citation Law.)
- Luguri, J., & Strahilevitz, L.J. (2021). "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006.


Sources:


- European Parliament, "New EU rules needed to address digital addiction," press release 20231208IPR15767. https://www.europarl.europa.eu/news/en/press-room/20231208IPR15767/new-eu-rules-needed-to-address-digital-addiction (vote 545/12/61, 12 December 2023; named features quoted verbatim).
- European Parliament, Texts adopted, "Addictive design of online services and consumer protection in the EU single market," TA-9-2023-0459. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0459_EN.html (primary text; retrieval of the full document body failed on , so only the press-release wording is relied on here).
- Office of the Texas Attorney General, press release on suit against Snap Inc. https://www.texasattorneygeneral.gov/news/releases/attorney-general-paxton-sues-snapchat-deceiving-parents-endangering-texas-kids-exposing-them (Snapstreaks named among features alleged to be addictive).
- Arkansas Advocate, "Arkansas files lawsuit against Snapchat parent company," 25 June 2026. https://arkansasadvocate.com/2026/06/25/arkansas-files-lawsuit-against-snapchat-parent-company/ (Snapstreaks named in the complaint).


Explicitly **not** cited: third-party gamification marketing blogs reporting Duolingo streak-freeze churn and retention percentages. These figures were seen during research, are not traceable to a Duolingo filing or paper, and are therefore treated as unverified and omitted.

D-05

# Public Metrics

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| D, Social and Identity | Habit | **Persuasive → Manipulative** | ●●●●● 5/5 |

## DEFINITION

A public metric is a count attached to a person or to a person's output, displayed to third parties, and controlled by the platform: followers, likes, views, reactions, upvotes, connection counts, reputation scores. The pattern is not the counting. Every product counts things. The pattern is the **publication of the count next to the identity**, which converts a private engagement signal into a public statement about the person's social worth. Once that conversion happens, the platform owns the scoreboard on which the user's self-image is kept. The user cannot leave without abandoning a score, cannot improve the score without producing more for the platform, and cannot appeal the score to anyone, because the platform sets the rules of the game and can change them unilaterally. You can identify it in the wild by asking one question: *is a number about this person visible to other people, and did the product decide to make it so?*

## PATTERN DNA

**Goal** raise content-creation volume and creator retention, which are the supply side of any social product → **Trigger** publication of a post, and thereafter every notification that the count has moved → **Interface** a number rendered adjacent to the user's name or content, updated in near-real-time, with a notification for each increment → **Bias** social proof, which supplies the interpretive frame (a high count means good); and social comparison, which supplies the pain (my count is lower than theirs) → **Emotion** validation on a good outcome, shame or inadequacy on a bad one, and, characteristically, *anticipatory* anxiety in the window between posting and the count settling → **Decision** post again; post differently; delete the underperforming post → **Behavior** increased content production, tuned toward whatever the count rewards → **Reward** the count itself, delivered on a **variable schedule**, because the user cannot predict which post will land → **Business metric** content supply, creator retention, session frequency (the user returns to check the number).

**The link that breaks on disclosure:** *Reward*. "This number is set by a ranking system we control, it is not a measurement of your worth, and we could change how it is computed tomorrow" is a true sentence which, if internalised, would substantially reduce the number's grip. The count's power depends on the user treating it as a measurement of something real.

## PSYCHOLOGICAL BASIS

1. **Social proof.** The display of others' behavior as evidence for what is good, and by extension for who is valued. This is one of Cialdini's **seven** principles (2021 expanded edition; Unity is the seventh, added after the original six). Social proof is among the better-supported of them, though per-principle replication varies and is not uniformly settled.

2. **Variable reinforcement.** The count's increments are unpredictable in timing and magnitude. This is a variable schedule, and variable schedules produce more persistent responding than reliable ones. The canonical primary source is **Ferster & Skinner (1957)**. The extension from pigeon key-pecking to human content-posting is an **analogy** drawn from the operant literature, not a finding within it, and is labelled as such. No response-rate figure is attached, because none was retrieved from the primary text.
3. **Loss framing.** A count that has gone up can go down (unfollows, deleted likes), and the value function is steeper for losses than for gains (Kahneman & Tversky 1979; replicated by Ruggeri et al. 2020 across 4,098 participants in 19 countries). A user with 10,000 followers is not experiencing a gain of 10,000; they are experiencing a permanent exposure to a possible loss.
4. **Social comparison.** The specific claim that public engagement counts drive upward social comparison and depress adolescent wellbeing is the central empirical question here and is **contested**. It is not resolved by any source retrieved for this document. CITATION NEEDED. What can be said without a citation is weaker and still useful: the platform that publishes the count is the party best placed to know its effects, and at least one such platform has acted as if the effect were real (see Instagram, below).

#### INTERFACE EXPRESSION

- A number rendered in the same visual unit as the identity: adjacent to the avatar, the handle, or the post.
- Real-time or near-real-time update, so the user has a reason to return specifically to watch the number.
- A notification per increment, which converts the metric into a trigger (A-03 Badge Signal, A-04 Notification Cascade).
- Ranked or leaderboard presentations, which convert the count from an absolute figure into a relative one (D-06 Comparative Ranking). This is the single largest escalation available in the pattern.
- Threshold effects: verified badges, monetisation tiers, and "top creator" markers gate benefits on the count, which converts the metric from a symbol into a currency (D-10 Status Marker).

#### REAL-WORLD EXAMPLES

**Instagram / Meta: the platform's own hiding of like counts.** On 26 May 2021 Meta announced that Instagram and Facebook users would be given the option to hide public like counts, both on other people's posts in their feed and on their own posts. In its own announcement, Meta stated that it had tested hiding like counts to see whether it might "depressurize people's experience on Instagram," and that feedback showed the change was beneficial to some people and annoying to others, particularly because people use like counts to gauge what is trending, which is why it was shipped as a *choice* rather than a default.

This is the most valuable single artifact in this entry, and it should be read carefully, because it cuts both ways.

- **It is evidence that the pattern has a cost.** A company does not build a feature to hide its own core engagement signal unless it believes the signal is doing damage. The word "depressurize" is Meta's, and it concedes that pressure exists.

- **It is also evidence of how little the concession costs.** The remedy shipped as an opt-in setting, not a default. Per I-01 Honest Default, an intervention that requires the harmed user to find and enable it is not a repayment of behavioral debt, because the users most affected by a status metric are the least likely to switch it off. frameworks/03 states the test: a remedy that costs the company nothing is not repaying anything. A default-off like count would have cost Meta something. An opt-in one does not.

**Meta: public metrics as alleged harm.** The bipartisan multistate attorneys-general action filed against Meta on 24 October 2023 alleges that Meta designed features that harm adolescent mental health, and identifies among them features operating through **social comparison**, specifically naming "likes" alongside photo filters, algorithmic ordering, and infinite scroll. These are allegations in pleadings and are unadjudicated. They are cited here as evidence that the pattern is contested at law, not as a finding of harm.

No description of any current product screen is given, per the evidence rule in PROJECT.md §7.

#### BUSINESS RATIONALE

Every social product has a two-sided market problem: consumption is easy to obtain and production is not. Public metrics solve the production side more cheaply than any alternative, because they pay creators in a currency the platform mints for free. A like costs the company nothing and is worth something to the recipient. No other compensation scheme has that margin.

The metric also produces three second-order benefits the company books but rarely names. It creates a **return trigger** (the user comes back to check the number). It creates a **switching cost**, because the count is not portable: leaving means starting from zero somewhere else (D-12 Reputation Lock-in). And it creates a **content-quality signal for free**, since the crowd's counting does the ranking work the platform would otherwise have to pay for.

A rational company ships this. It is not merely profitable, it is close to structurally necessary for a social product to exist at all. That is exactly why the ethical analysis has to be careful: "do not publish counts" is not a serious proposal, and a document that made it would be ignored.

#### BEHAVIORAL COST INDEX

Scored per frameworks/02. Two implementations.

| AXIS | PRIVATE-BY-DEFAULT METRIC (CREATOR SEES OWN COUNT; PUBLIC DISPLAY IS OPT-IN) | PUBLIC COUNT NEXT TO IDENTITY, NOTIFIED PER INCREMENT, TIED TO STATUS TIERS |
|---|---|---|
| Time | 1 | 3 |
| Attention | 1 | 4 |
| Memory | 0 | 1 |
| Trust | 1 | 3 |
| Money | 0 | 1 |
| Privacy | 1 | 3 |
| **Emotion** | **2** | **5** |

**Left column.** *Time 1 / Attention 1:* the creator can see performance data when they choose to look. *Emotion 2:* seeing your own numbers can still sting, but the user is aware and has chosen to look; mild friction. **Max = 2 → Ethical.** Note that this column describes a *real* product configuration, not a fantasy: analytics dashboards for creators, with no public display, are an existing and viable design.

**Right column.** *Attention 4:* the count generates a notification stream the user did not budget for, and the return-to-check behavior is not consciously chosen. *Trust 3:* the user is not fully aware that the number is a platform-controlled output, not a measurement. *Privacy 3:* the user's popularity, and its trajectory, are disclosed to everyone, which most users would not choose to publish as a standalone fact. *Money 1:* only where status gates monetisation. *Emotion 5:* **the load-bearing score.** The design ties self-worth to a number the product controls, the user is not aware that the number is manufactured rather than measured, awareness would change their relationship to it, **and the pattern works precisely because of that unawareness**. A user who fully believed "this number is an artifact of a ranking system, and it says nothing about me" would not check it fifteen times an hour. **Max = 5 → Manipulative.**

**A discrepancy with TAXONOMY.md, recorded rather than smoothed over.** The taxonomy gives D-05 a valence *range* of Persuasive → Manipulative, but the left-hand column above scores a maximum of 2, which under the BCI's derivation rule (frameworks/02: Ethical = every axis ≤ 2) is **Ethical**. The instrument says the declared floor of the range is one band too high, and the same is true of B-09. Both should be reconciled when the taxonomy's ranges are re-derived from the BCI rather than from judgment, which TAXONOMY.md open question 4 already identifies as necessary.

That Emotion 5 is what justifies the severity ceiling of 5 in TAXONOMY.md, and it is the same structural argument made for C-01: the product manufactures an affect and then depends on the user not knowing it was manufactured.

#### VALENCE ANALYSIS

**The case for Ethical, which exists and is usually skipped.** Feedback is not manipulation. A writer who cannot tell whether anyone read the piece is worse off than one who can. Public counts also do genuine informational work for *readers*: they are a cheap quality signal in an environment with no other, and Meta's own finding, that removing like counts annoyed users because they use counts to gauge what is trending, is direct evidence that the metric serves consumers and not only the platform. Any analysis that treats public metrics as pure extraction has to explain that finding, and cannot.

**The case for Persuasive.** Publishing the count is a design choice that recruits the user's self-image into the service of the platform's content supply. The user knows the count is public. They know others can see it. There is no deception at the level of fact. What there is, is a **conflict of interest that is never disclosed**: the platform benefits when the user cares about the number, and every design decision around the number (notify on increment, render it large, rank by it) is taken by a party that profits from the user caring more.

**The case for Manipulative, and where the line falls.** The pattern crosses into manipulation at the point where the number stops being *information about the content* and becomes *a statement about the person*. Three markers:

1. **The count is attached to the identity rather than to the artifact.** A view count on a video is information. A follower count under a name is a rank.
2. **The count is notified rather than looked up.** A metric the user must go and check is a tool. A metric that arrives unbidden, incrementally, on a variable schedule, is a reinforcement device.
3. **The count is not portable.** A score that cannot leave the platform is not the user's; it is the platform's, held in the user's name (D-12).

Under the Disclosure Test, the chain snaps at *Reward*. The number's grip depends on the user experiencing it as a measurement. Tell them it is a manufactured output of a system optimised for the platform's content supply, tell them the ranking could change tomorrow and the number with it, and the affect the pattern depends on does not survive. That is the structural definition of manipulation used throughout this document, and public metrics meet it.

**The honest complication.** Unlike C-01, this pattern **cannot simply be removed**, because part of its function (informing readers, rewarding creators) is legitimate and load-bearing. The remedy is therefore necessarily a matter of degree, and the analysis has to live with that. This is a case where the correct output is not a prohibition but a set of design constraints, below.

#### ETHICAL ALTERNATIVE

The legitimate goal, rewarding creators and signalling quality to readers, survives every one of these changes. What does not survive is the pressure.

1. **Split the audience for the number.** Show the creator their own metrics in full. Do not show them to anyone else by default. This preserves the feedback function entirely and eliminates the ranking function.
2. **If counts must be public, make the private view the default and the public view the opt-in**, not the reverse (I-01 Honest Default). Meta shipped the opposite polarity in 2021, and the polarity is the whole argument.
3. **Do not notify on increment.** A metric the user retrieves is a tool. A metric that pushes itself is a slot machine. Removing the per-like notification forfeits sessions, which is what makes it a real concession.
4. **Aggregate and delay.** Report engagement in daily buckets rather than in real time. This preserves all the information and destroys the variable-reinforcement schedule, because a scheduled report is a fixed-interval one.
5. **Make the score portable** (I-11 Portability by Design). A creator who can export an attested audience and reputation is a creator whose standing is theirs. This directly forfeits lock-in, which is precisely why almost nobody does it, and why doing it would be a credible signal.
6. **Publish the fact that the number is manufactured.** State, in the product, that engagement counts are outputs of a ranking system that the platform controls and can change. This is I-04 Usage Transparency applied to social metrics, and it is the one intervention that attacks the Emotion 5 directly.

#### REGULATORY STATUS

- **No jurisdiction regulates public engagement metrics as such.**
- **United States.** Public metrics appear as *alleged conduct* in the bipartisan multistate attorneys-general litigation against Meta filed 24 October 2023 (33 states in federal court, with additional states filing in their own courts), which names "likes" among the features alleged to harm adolescents through social comparison. Brought under state consumer-protection statutes and COPPA. **Unadjudicated.**
- **European Union.** The European Parliament's resolution of 12 December 2023 on addictive design (adopted 545/12/61) does not name public metrics, though it names neighbouring mechanisms (infinite scroll, autoplay, constant push and read-receipt notifications) and asks the Commission to consider banning addictive techniques not covered by the Unfair Commercial Practices Directive. Public metrics are one of the more obvious gaps in that list.
- **Fake metrics are a different matter and are already regulated.** Where the count is *fabricated* rather than merely published, it becomes "fake social proof" in Brignull's taxonomy and falls within existing deception law. Mathur et al. (2019) measured social-proof deception at scale, finding "Activity Messages" and "Testimonials of Uncertain Origin" among 15 dark-pattern types across ~11,000 shopping sites. Note the sharp asymmetry: a **false** count is illegal almost everywhere, and a **true** count that destroys a teenager's self-image is legal almost everywhere. That asymmetry is a direct consequence of organising law around deception, and it is one of this document's central arguments.

#### RELATED PATTERNS

- **D-01** Social Proof Counter, the same number read as a signal about the *content* rather than the *person*. The distinction is the whole entry.
- **D-06** Comparative Ranking, the escalation from absolute count to relative position.
- **D-07** Identity Investment and **D-12** Reputation Lock-in, the switching costs the count creates.
- **D-10** Status Marker, where the count is converted into a conferred distinction.
- **C-02** Variable Reward, the reinforcement schedule the notification stream runs on.
- **A-03** Badge Signal, **A-04** Notification Cascade, the delivery mechanisms.
- **I-01** Honest Default, **I-04** Usage Transparency, **I-11** Portability by Design: the ethical counterparts.

#### SOURCES

Verified project sources (see research/sources/):


- Cialdini, R.B. (2021). *Influence, New and Expanded: The Psychology of Persuasion*. New York: Harper Business. ISBN 9780062937650. (Social proof; **seven** principles, per influenceatwork.com. Per-principle replication varies.)
- Ferster, C.B., & Skinner, B.F. (1957). *Schedules of Reinforcement*. New York: Appleton-Century-Crofts. (Book-level citation, no figures attached. The application to social feedback is labelled an analogy.)
- Kahneman, D., & Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2), 263-291. DOI 10.2307/1914185; Ruggeri, K., et al. (2020). *Nature Human*

*Behaviour* 4(6), 622-633. DOI 10.1038/s41562-020-0886-x.

- Mathur, A., et al. (2019). "Dark Patterns at Scale." *PACM HCI* 3(CSCW), Art. 81. DOI 10.1145/3359183. (~11,000 shopping sites; 1,818 instances; 15 types / 7 categories, including a Social Proof category.)
- Brignull, H., deceptive.design /types (18 types, including Fake social proof).

Sources:

- Meta, "Giving People More Control on Instagram and Facebook," 26 May 2021. https://about.fb.com/news/2021/05/giving-people-more-control/ (first-party announcement of the option to hide like counts; "depressurize" is Meta's own word).
- NBC News, "Facebook and Instagram to let users hide like counts," 26 May 2021. https://www.nbcnews.com/tech/social-media/instagram-rolls-out-option-users-hide-likes-n1268624 (contemporaneous reporting; the finding that some users wanted counts retained in order to see what was trending).
- CNBC, "Meta sued by 42 attorneys general alleging Facebook, Instagram features are addictive and target kids," 24 October 2023. https://www.cnbc.com/2023/10/24/bipartisan-group-of-ags-sue-meta-for-addictive-features.html ("likes" named among the alleged social-comparison features).
- Office of the New York Attorney General, press release on the multistate Meta suit, October 2023. https://ag.ny.gov/press-release/2023/attorney-general-james-and-multistate-coalition-sue-meta-harming-youth

**Marked as unresolved:** the empirical question of whether public engagement counts causally depress adolescent wellbeing. CITATION NEEDED. No source retrieved for this document settles it, and the entry deliberately does not assert it.

E-03

# Contact Harvesting

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| E, Growth and Acquisition | Expansion | **Persuasive → Manipulative** | ●●●●● 5/5 |

## DEFINITION

Contact harvesting is the transfer of a user's address book, or an equivalent external contact list (email account contacts, phone contacts, messaging graph), from the user's device or account to the company's servers, where it is retained and used for purposes beyond the immediate action the user requested. The pattern's defining property is a **consent mismatch across parties**: the user consents, and the *contacts*, who never met the product, do not. Whatever the user agreed to, the third parties in that address book, including people who are not users, may not be, and cannot be, asked. The company thereby acquires an asset (a social graph) that was never theirs and was never the user's to give. Identifying it in the wild: any flow in which a "find your friends" affordance results in server-side storage of names, numbers, or addresses belonging to people who have no relationship with the product.

The mechanism should be kept distinct from the *ask* (B-11 Contact Ingestion Prompt) and from the *outbound messaging* it enables (E-04 Invite Impersonation). This entry is about the taking and the keeping.

## PATTERN DNA

**Goal** acquire new users at near-zero marginal cost by mining the existing user's relationships → **Trigger** an onboarding step framed as a benefit ("find people you know"), placed at the point of maximum goodwill (B-09, B-11) → **Interface** a single affirmative button, an OS permission dialog, and, in the manipulative form, a pre-checked or absent selection step so that *all* contacts are uploaded rather than the ones the user meant → **Bias** the user is evaluating a *benefit to themselves* (finding friends), not a *disclosure about others*, so the party whose interests are at stake is not represented in the decision at all → **Emotion** sociability, mild curiosity, no threat perception whatsoever → **Decision** "yes, connect me" → **Behavior** the full address book is transmitted → **Reward** a list of people the user recognises, delivered immediately, which validates the choice → **Business metric** viral coefficient, invite-send volume, CAC, and, less visibly, the size and quality of the shadow graph the company now holds.

**The link that breaks on disclosure:** *Interface*, and unusually the disclosure that breaks it is a disclosure **to a party who is not present**. "Your friend Sarah has uploaded your name, phone number, birthday, and email address to our servers, where we will retain them and use them to model your social graph, whether or not you ever use this product" is the sentence. It is true, it is never shown to Sarah, and it cannot be, because Sarah is not in the room. That structural impossibility is what makes this pattern distinctive and is why its severity ceiling is 5.

## PSYCHOLOGICAL BASIS

1. **The consenting party is not the affected party.** This is the core, and it is not a cognitive bias, it is a structural defect in the consent architecture. Every framework in the behavioral literature, from choice architecture (Thaler & Sunstein) to Fogg's kairos, models a single decision-maker weighing costs and benefits to themselves. None of them models the case where the cost falls on someone outside the interaction. Consent theory has no way to price an externality, and neither does the interface. **This is offered as an original observation of this document, not as a finding from a cited source.**
2. **Timing (kairos).** Fogg (2009) proposes that behavior is a product of motivation, ability, and triggers, and introduces *kairos*, the opportune moment to persuade. The address-book ask is placed at the moment of highest goodwill and lowest scrutiny: immediately after signup, before any negative experience has occurred, framed as part of getting started. Note that Fogg's 2009 paper contains **no equation**; "B=MAT" is a later community retrofit and is not used here.
3. **Reciprocity and social framing.** The request is framed socially ("see who's already here"), which recruits Cialdini's reciprocity and liking principles rather than presenting the transaction as what it is, a bulk data transfer. Cialdini's principles number **seven** in the 2021 expanded edition; per-principle replication varies and is not presented here as settled.
4. **Default and scope blindness.** Users evaluating "share my contacts" do not spontaneously enumerate what is *in* their contacts: doctors, lawyers, ex-partners, sources, sponsors, employers. The gap between the imagined scope of an address book and its actual scope is, in this document's view, the largest single unmeasured disclosure in consumer software. CITATION NEEDED for any empirical measurement of that gap.

### INTERFACE EXPRESSION

- A "Find friends" or "Connect your contacts" card in onboarding, styled as a feature rather than a permission.
- A soft pre-prompt (B-09) preceding the OS dialog, protecting the one-shot system ask.
- **No selection step.** The user is not shown the list and asked which contacts to use. The absence of that screen is the pattern.
- Upload of the *entire* address book, including fields the stated purpose does not require (birthdays, physical addresses, notes).
- **Retention.** The contacts are kept after the immediate purpose is served, which is where the pattern stops being a feature and becomes an asset.
- In the aggressive form, upload occurs even when the user did *not* select the option, which is no longer a design question but a misrepresentation (see Path, below).

### REAL-WORLD EXAMPLES

**Path, Inc. (2013), FTC enforcement.** The Federal Trade Commission's action against the social-networking app Path is the foundational documented case, and it is the FTC's **first public enforcement action against a mobile app addressing the collection and use of a device user's address-book contacts**. Per the FTC's press release, Path automatically collected and stored personal information from the user's address book **even when the user did not select the "find friends from your contacts" option**, capturing first and last names, addresses, phone numbers, email addresses,

Facebook and Twitter usernames, and dates of birth for each contact. Path paid **$800,000** to settle charges of illegally collecting personal information from children without parental consent, was required to establish a comprehensive privacy programme and to obtain independent privacy assessments every other year for twenty years, and deleted the address-book information collected during the period at issue.

Two features of this case are worth naming precisely. First, the fine attached to the **COPPA** violation (children's data), not to the address-book collection as such, which tells you where the enforceable law actually was. Second, the conduct was uploading **without the user selecting the option**, which is deception, and therefore reachable. A product that had asked, and been told yes, and then uploaded everything, would have been on far safer legal ground while extracting substantially the same asset.

**LinkedIn, *Perkins v. LinkedIn Corp.* (N.D. Cal.).** LinkedIn's "Add Connections" feature invited members to import contacts from external email accounts, and generated invitation emails to those contacts; where no response came, LinkedIn automatically sent two further reminder messages. The class action alleged that although the messages appeared to be endorsements from the member, users did not compose them, did not consent to multiple messages being sent on their behalf, and were not compensated for the use of their name or likeness. LinkedIn settled for **$13 million**, covering a class of approximately **20.8 million members**, for the period **17 September 2011 to 31 October 2014**, and made disclosure and functionality changes, including clarifying that up to two reminders are sent per invitation and, by the end of 2015, allowing members to stop reminders by cancelling the invitation.

Note what this case is really about: the harvesting was upstream, the litigable harm was downstream (the impersonated messages, E-04). This is the recurring shape. **The taking is rarely what gets sued; the sending is.**

No description of any current product's onboarding screen is offered, per the evidence rule in PROJECT.md §7.

#### BUSINESS RATIONALE

Contact harvesting is the cheapest customer acquisition channel that has ever existed. The cost of acquiring a targeted, pre-qualified, socially-vouched lead through paid media is real money. The cost of acquiring the same lead by asking an existing user to hand over their address book is one modal.

The asset is also durable in a way that paid acquisition is not. An uploaded address book yields three distinct assets, only the first of which the user is told about:

1. **Immediate friend-matching**, which is the stated benefit and is genuine.
2. **An invite target list**, which powers outbound growth and, in the aggressive form, E-04.
3. **A shadow social graph**, including edges to and among people who are not users and have never consented to anything. This third asset is the most valuable and the least disclosed. It permits people-you-may-know inference, ad targeting, and the reconstruction of relationships that neither party disclosed.

A rational growth team ships this. The viral coefficient is the single most leveraged number in a consumer-social P&L, and this is the cheapest input to it.

BEHAVIORAL COST INDEX

Scored per frameworks/02. Note that the BCI, like every consent framework, is built to score **the user**, and this pattern's principal victim is not the user. That limitation is stated rather than papered over, and a second scoring is given for the third party.

**Scored for the consenting user:**

| AXIS | SELECTIVE, TRANSIENT MATCHING (USER PICKS CONTACTS; NOTHING RETAINED) | FULL UPLOAD, RETAINED, USED TO BUILD A GRAPH |
|---|---|---|
| Time | 1 | 1 |
| Attention | 1 | 2 |
| Memory | 0 | 1 |
| Trust | 1 | 4 |
| Money | 0 | 0 |
| **Privacy** | **2** | **5** |
| Emotion | 0 | 3 |

*Left column:* the user chose specific contacts, the match was performed, nothing was kept. Costs exist and are visible. **Max = 2 → Ethical.** *Right column:* **Privacy 5** is the load-bearing score, and it is a 5 rather than a 4 because the design **works precisely because of the unawareness**: a user who understood that the full book, including fields irrelevant to matching, would be retained indefinitely and used to model non-users, would frequently decline, and the growth loop depends on them not understanding that. *Trust 4:* the user would feel betrayed on learning, and awareness would change behavior. *Emotion 3:* shame and social exposure follow, but only later, when the invites arrive. **Max = 5 → Manipulative.**

**Scored for the harvested third party** (the person in the address book who is not a user):

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 0-2 | Zero, unless they receive invitations and reminders, which is time taken from someone who never engaged with the product at all. |
| Attention | 2 | Unsolicited contact. |
| Memory | 0 | None. |
| Trust | 5 | They have no relationship with the company to spend, so what is damaged is trust in the *friend* who handed them over. The harm is displaced onto a relationship the company does not own. |
| Money | 0 | None. |
| **Privacy** | **5** | Their name, number, email, and birthday are on a server belonging to a company they have never heard of. They were never asked. They cannot be asked. They cannot decline. |
| Emotion | 2 | Irritation, and in adversarial cases (a source, a stalking victim, a person with a secret) far worse. |

**Max = 5 → Manipulative**, and note that for the third party there is **no ethical column at all**. Even the benign implementation exposes them. That is the finding: contact harvesting is the one pattern in this taxonomy for which the affected party has no version of the design that protects them, short of the data never leaving the device.

#### VALENCE ANALYSIS

**The case for Persuasive, honestly stated.** The stated benefit is real. Social products are useless without a social graph, and a new user with no connections is a user who will churn. Contact matching solves a genuine cold-start problem, it delivers immediate value, and users like it: given a clear choice, many would say yes. There is nothing intrinsically deceptive about asking a user whether they would like to find people they know. If the ask is honest, the scope is minimal, and nothing is retained, the pattern is not merely persuasive, it is **ethical**, and the left-hand column above says so.

**The case for Manipulative.** It rests on three claims, and the third is the one that matters.

1. **The scope exceeds the stated purpose.** Matching requires a hash comparison. It does not require birthdays, physical addresses, or indefinite retention. Every field taken beyond the purpose is a cost paid without being priced.
2. **The retention is undisclosed and is the actual point.** The user consents to a *lookup* and funds an *asset*. That gap is a Privacy 5 under the BCI: unaware, would change behavior, and the design depends on the unawareness.
3. **The consent is structurally void for the party who bears most of the cost.** This is the argument that makes E-03 different in kind from every other pattern in this document. In G-01, the user's consent is *degraded*. Here, the affected party's consent is *impossible*. The person in the address book cannot decline, cannot be informed, cannot opt out, and in most cases will never learn that it

happened. No amount of interface improvement fixes this, because the interface is not shown to them. **A pattern whose harm cannot be cured by any possible disclosure to the harmed party is manipulative by construction**, and that is the strongest form of the Disclosure Test in this document.

**The uncomfortable corollary.** It follows that even a scrupulously honest, fully-disclosed, opt-in, user-loves-it contact upload still harvests people who did not consent. The correct verdict is therefore not "do it honestly" but "do it **without transmitting the data**," which is a technical constraint rather than a copy change, and which is what the ethical alternative below actually demands. This is the only entry among these nine where the remedy is architectural rather than communicative.

#### ETHICAL ALTERNATIVE

The legitimate goal, connecting a new user to people they already know, is real and can be achieved without acquiring anyone's address book.

1. **Match on-device.** Compute the intersection locally: the device holds the contacts, the server supplies a bloom filter or a private-set-intersection protocol, and only the *matches* are surfaced. The company learns who matched. It does not learn who did not. This is the whole remedy, and it is technically routine.
2. **If a server-side lookup is unavoidable, transmit salted hashes, retain nothing, and say so.** Non-matching hashes are discarded on the same request. Retention is the pattern; a lookup is not.
3. **Show the list. Let the user choose.** A selection step, defaulted to nothing selected, converts a bulk transfer into a set of individual decisions. It will reduce upload volume by a large margin, which is the point: the reduction *is* the consent that was previously missing.
4. **Take only the fields the purpose requires.** A phone number matches. A birthday does not.
5. **Give the non-user a right that survives.** Any third party should be able to query and delete their own presence in the company's contact store without creating an account. Almost no product offers this. Offering it is expensive and forfeits the shadow graph, which is exactly what makes it a genuine repayment of behavioral debt (frameworks/03) rather than a policy page.
6. **Never send on the user's behalf without a per-message confirmation** (see E-04, and *Perkins v. LinkedIn*).

#### REGULATORY STATUS

- **United States.** *In the Matter of Path, Inc.* (FTC, 2013): $800,000 civil penalty and a twenty-year privacy-assessment obligation, for collecting address-book data without consent and, specifically, collecting children's personal information in violation of **COPPA**. The FTC describes it as its first public enforcement action against a mobile app on address-book collection. Note again the legal shape: the penalty rode on COPPA and on deception (collecting when the user had *not* opted in), not on address-book collection per se.
- **United States, private litigation.** *Perkins v. LinkedIn Corp.* (N.D. Cal.): $13 million class settlement, class of ~20.8 million members, covering 17 September 2011 to 31 October 2014, over invitation and reminder emails generated from imported contacts and appearing to come from the member. Claims sounded in right-of-publicity and consent, not in data protection.

- **European Union.** Uploading an address book is processing the personal data of the contacts, who are data subjects with no relationship to the controller. GDPR's requirements for a lawful basis, purpose limitation, and data-subject rights apply to *them*, and this is very difficult for a controller to satisfy, since the contacts cannot practically be informed. CITATION NEEDED for the specific GDPR articles and for any enforcement decision on contact upload; neither was retrieved in this research.
- **Platform rules.** Apple's App Store Review Guideline 5.1.1(iv), retrieved verbatim: "*Apps must respect the user's permission settings and not attempt to manipulate, trick, or force people to consent to unnecessary data access.*" Guideline 5.1.2(i) forbids requiring users to enable system functionalities in order to access functionality, content, or compensation. Platform review is, in practice, a faster-acting constraint on this pattern than any statute.

### RELATED PATTERNS

- **B-11** Contact Ingestion Prompt, the ask. **B-09** Permission Priming, the pre-prompt that protects it.
- **E-04** Invite Impersonation, the outbound abuse this enables, and the thing that actually gets litigated (*Perkins*).
- **E-02** Viral Invite Loop, **E-07** Network Pressure, the growth machinery downstream.
- **G-10** Inferred Profile, the shadow graph built from non-users' data. E-03 is the largest single input to G-10 in consumer software.
- **G-04** Bundled Consent, **G-03** Default Over-disclosure, the adjacent consent failures.
- **I-11** Portability by Design is *not* the counterpart here. There is no Family I entry that cures E-03, because the cure is architectural (on-device matching), not a design pattern. **This is a gap in Family I and is recorded as such.**

### SOURCES

Verified project sources (see research/sources/):


- Fogg, B.J. (2009). "A Behavior Model for Persuasive Design." Persuasive '09, Article 40. DOI 10.1145/1541948.1541999. (Motivation, ability, triggers; *kairos*. The paper contains **no equation**.)
- Cialdini, R.B. (2021). *Influence, New and Expanded*. New York: Harper Business. (Seven principles; replication varies by principle.)
- Thaler, R.H., & Sunstein, C.R. (2021). *Nudge: The Final Edition*. New Haven: Yale University Press. (Choice architecture as a framework; no effect size asserted.)
- Gray, C.M., et al. (2018). "The Dark (Patterns) Side of UX Design." CHI '18. DOI 10.1145/3173574.3174108. (Reproduces Brignull's typology as Table 1, which includes **Friend Spam** and **Privacy Zuckering**, the two inherited names closest to this pattern.)
- Luguri, J., & Strahilevitz, L.J. (2021). "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006.


Sources:


- Federal Trade Commission, "Path Social Networking App Settles FTC Charges it Deceived Consumers and Improperly Collected Personal Information from Users' Mobile Address Books,"

February 2013. https://www.ftc.gov/news-events/news/press-releases/2013/02/path-social-networking-app-settles-ftc-charges-it-deceived-consumers-improperly-collected-personal ($800,000; first FTC mobile-app address-book action; collection even when the option was not selected; 20-year assessment obligation).

- Federal Trade Commission business-guidance blog, "FTC Path case helps app developers stay on the right, er, path," February 2013. https://www.ftc.gov/business-guidance/blog/2013/02/ftc-path-case-helps-app-developers-stay-right-er-path
- *Perkins v. LinkedIn Corp.*, No. 13-CV-04303-LHK (N.D. Cal.), settlement materials. https://www.casemine.com/judgement/us/627d6df8714d585d57ca6250 and https://topclassactions.com/lawsuit-settlements/closed-settlements/linkedin-add-connections-class-action-settlement/ ($13m; ~20.8m class members; 17 Sep 2011 to 31 Oct 2014; two automated reminders per invitation).
- Apple, *App Store Review Guidelines*, 5.1.1(iv) and 5.1.2(i). https://developer.apple.com/app-store/review/guidelines/ (quoted verbatim).

F-09

# Roach Motel

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| F, Monetization and Pricing | Exit | Manipulative | ●●●●● 5/5 |

**Attribution:** The name is **not original to this work**. It is Harry Brignull's, from the deceptive-design pattern library, and is reproduced in **Gray et al. (2018), Table 1**, which quotes Brignull's typology of eleven types, "Roach Motel" among them. This attribution is **verified** (ATTR VERIFIED): Gray et al. (2018) was retrieved and read in full, and the table was confirmed from the PDF. On Brignull's own current site the equivalent published type is **"Hard to cancel"**, one of eighteen. We retain "Roach Motel" as the mechanism name because it describes the *asymmetry* rather than only the cancellation case, and we credit Brignull.

## DEFINITION

A roach motel is a product structure in which the cost of entry and the cost of exit are deliberately made asymmetric: entry is engineered to be as frictionless as the company can make it, and exit is engineered to be as costly as the company can make it without triggering enforcement. The pattern is **not** the friction on exit. Friction on exit is sometimes legitimate (a confirmation before deleting a year of work is friction that serves the user). The pattern is the **asymmetry**, because the asymmetry is the only thing that reveals the intent. A company that believes its own product is worth having has no reason to make leaving harder than arriving; a company that makes leaving nine times harder than arriving has told you, in the only language a product has, what it thinks of its own retention.

Identifying it in the wild requires exactly one measurement, and it is the cheapest diagnostic in this document:

**The Asymmetry Ratio** = *(steps to start) ÷ (steps to stop) ≈ 1: symmetric. 2 to 3: mild retention design.* **> 3: the product is structurally opposed to your leaving.** *(Behavioral Gravity, frameworks/03. Original framework; proposed here.)*

## PATTERN DNA

**Goal** suppress voluntary churn, and specifically suppress churn among users whose intent to leave is real but whose determination is finite → **Trigger** the user's own decision to leave, which is the only trigger in this taxonomy supplied entirely by the user → **Interface** an entry path of one or two clicks, and an exit path of multiple screens, interstitial offers, re-confirmations, channel changes (web to phone, app to web), and hours-of-operation constraints → **Bias** the sunk cost of the effort already expended in the exit flow itself; present bias, since abandoning the cancellation is the path of least immediate effort; and simple depletion, because the flow is designed to outlast the user's motivation → **Emotion** irritation, then fatigue, then resignation → **Decision** "I will do this later," which is the decision the design is engineered to produce → **Behavior** abandonment of the cancellation attempt, and

payment of at least one further billing cycle → **Reward** none for the user; this is the rare pattern with **no user-side reward at any link** → **Business metric** churn rate, retained MRR, LTV.

**The link that breaks on disclosure:** *Interface*. The pattern's entire function is to convert a user's decision into a user's fatigue, and the fatigue depends on the user not knowing how many screens remain. Print "this cancellation takes 4 pages, 6 clicks, and 15 options, and you may stop at any time and we will still cancel you" at the top of the flow, and the flow stops working. Note the diagnostic significance: **F-09 is the only pattern among these nine whose DNA chain contains no user reward whatsoever.** That absence is why its valence range in TAXONOMY.md has no ethical or persuasive band. There is nothing to argue about.

#### PSYCHOLOGICAL BASIS

1. **Sludge.** Thaler and Sunstein's own term for friction that keeps people from what they *do* want. This is the adversarial inverse of a nudge, and it is the correct primary frame for F-09. Cite the concept as a framework, **not** as an effect size: the pooled nudge effect is disputed under publication-bias correction and no figure is asserted here.
2. **Obstruction, empirically.** Luguri & Strahilevitz (2021) is the strongest available evidence and it is unusually direct. In two randomized experiments on representative samples of American consumers (Study 1 n = 1,963; Study 2 n = 3,777), acceptance of a dubious service rose from **11.3%** in the control condition to **25.8%** under mild deceptive patterns and **41.9%** under aggressive ones (the paper renders these as a 228% and a 371% increase). Study 2 identified **obstruction** among the three strategies "particularly likely to manipulate consumers successfully," alongside hidden information and trick questions. Most damningly for the standard economic defence of these designs, the authors found that under deceptive patterns **the price became immaterial**: "Decision architecture, not price, drove consumer purchasing decisions."

That finding is the empirical backbone of this entry. It disposes of the argument that a user who really wanted to leave would simply leave. The experiment shows that architecture beats preference.

1. **Present bias.** Abandoning the exit flow is immediately cheaper than completing it, and the cost of abandonment (another billing cycle) is deferred. CITATION NEEDED for a specific present-bias citation; none was retrieved for this document.

#### INTERFACE EXPRESSION

- **Entry:** one-click enrol, often with the enrolment offered as a side-effect of an unrelated action (a checkout, a shipping-speed selection). See F-13 Basket Sneaking and F-12 Preselected Add-on for the mechanics.
- **Exit:** a sequence, and the sequence has characteristic components:
- Re-confirmation loops. The user presses a button labelled to end the relationship, and the relationship does not end.
- Interstitial retention offers (H-02) that are not skippable.
- Guilt copy on the decline path (F-08 Confirmshaming).
- Benefit reminders, listing what will be lost.

- Channel changes: cancel online, but only after logging in on the web, but the account was created in the app.
- Human gatekeeping: phone-only cancellation, business hours only (H-11 Support Obstruction).
- **The tell:** a button that says "Cancel" and does not cancel.

#### REAL-WORLD EXAMPLES

**Amazon Prime, FTC v. Amazon.com, Inc. (ROSCA).** This is the best-documented instance in existence, because it was litigated to the fourth day of a jury trial before settling, and the internal name for the cancellation flow entered the public record.

Documented facts, from the FTC's own announcement and contemporaneous reporting of the trial:

- On **25 September 2025** Amazon settled with the FTC for **$2.5 billion**: a **$1 billion civil penalty** and **$1.5 billion in refunds** to consumers. The FTC describes it as among the largest settlements in the agency's history. The settlement was reached on the fourth day of trial; jury proceedings had begun on 22 September 2025.
- The claims were brought under **Section 5(a) of the FTC Act** and the **Restore Online Shoppers' Confidence Act (ROSCA)**.
- Amazon's internal name for the Prime cancellation flow was the **"Iliad Flow."** Trial reporting describes it as a **four-page, six-click, fifteen-option** sequence containing repeated diversions, including discount offers and reminders of Prime benefits. **Enrolment could be completed in two clicks.**
- The court found a dispute of material fact as to whether Amazon offered a simple cancellation mechanism, because pressing the **"End Membership"** button **did not end the membership**: customers were required to reaffirm the intention to cancel three times before cancellation took effect.
- Per the FTC's counsel, by Amazon's own estimation there were more than **35 million non-consensual Prime enrolments** over a seven-year period.
- The settlement requires Amazon to present a clear button to cancel and to provide a simple cancellation mechanism **through the same medium the customer used to enrol**.

Apply the diagnostic. Steps to start: **2**. Steps to stop: **6 clicks across 4 pages**, with 3 re-affirmations. **Asymmetry Ratio ≈ 3, and by page count ≈ 3 to 4.** The Behavioral Gravity threshold for "structurally opposed to your leaving" is > 3. The measurement and the $2.5 billion settlement agree, which is the closest thing this document has to an external validation of one of its own instruments.

The naming is itself evidence, and it is worth dwelling on. "Iliad" is a reference to a long, arduous epic. **The company named the flow after its own difficulty.** The intent was not inferred by a regulator from a pattern of clicks; it was written down internally, in a joke.

No other product's cancellation flow is described here from memory. Cancellation flows change, frequently in response to exactly this kind of enforcement, and a remembered flow is not evidence about the present one (PROJECT.md §7).

#### BUSINESS RATIONALE

The arithmetic is brutally simple, and it is why this pattern will not go away on its own. A subscription business's LTV is a function of churn. A cancellation flow that causes some fraction of intending cancellers to give up converts each of them into at least one more billing cycle, and often into indefinite continuation, because a user who abandoned the flow once is unlikely to attempt it again soon. The cost of building the flow is one sprint. The revenue is recurring, compounding, and immediately visible in the retention dashboard.

Worse, the pattern is **self-validating**. Churn goes down after the flow ships. The dashboard reports this as a retention win. Nothing in the instrumentation distinguishes "users decided to stay" from "users failed to leave," and no standard analytics stack even collects the variable that would separate them (attempted-cancellation starts versus completions). This is Behavioral Debt (frameworks/03) in its clearest form: the metric books the interest payment and reports it as income.

The Amazon settlement is the counter-argument, and it is the only counter-argument that has ever moved this pattern: **$2.5 billion is a number large enough to appear on the same balance sheet as the retained revenue.** Until enforcement is priced in, a rational company ships the roach motel. That is not cynicism, it is the actual decision structure, and any ethics argument that ignores it is not going to be persuasive to the people who make the decision.

#### BEHAVIORAL COST INDEX

Scored per frameworks/02. **There is only one column, and that is the finding.** Unlike C-01, A-01, B-09, D-05, or E-03, this pattern has no implementation in which the costs are visible and accepted, because if the costs were visible and accepted the pattern would not function. The absence of an ethical column is not an oversight; it is the argument.

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 4 | Minutes to hours, sometimes across multiple attempts and channels. The user did not budget for it, would not have accepted it if told in advance, and the length is concealed until it has been incurred. |
| Attention | 3 | Sustained vigilance is required to avoid being diverted onto an offer, a pause, or a downgrade the user did not want. |
| Memory | 3 | The user must hold in mind that they have *not yet* cancelled, and must remember to verify. Products in this class routinely accept a "cancel" instruction and continue billing, which is the specific harm this axis measures. |
| Trust | 5 | The betrayal is total and legible: the user learns, unmistakably, that the company built a machine to defeat them. Nothing else in a product communicates contempt so clearly, and no user forgets it. |
| **Money** | **5** | **The load-bearing score.** The user pays for a service they have decided to stop using, having taken an action they believed ended it. They are unaware they are still being charged, awareness would immediately change their behavior, **and the entire design exists in order to produce that unawareness.** By Amazon's own estimate, per the FTC, this reached 35 million non-consensual enrolments. |
| Privacy | 0 | None. This pattern takes money and time, not data. |
| Emotion | 3 | Frustration and humiliation, deliberately induced, but the user is aware of them. |

**Max = 5 → Manipulative.** No implementation of a roach motel scores below 4, because a roach motel that scored below 4 would be a symmetric exit, which is a different pattern with a different ID (**I-03**).

### VALENCE ANALYSIS

This is the only entry among these nine where the valence is not genuinely contested, and it is worth being explicit about **why**, because a document whose thesis is "valence is a property of the implementation" owes an account of the patterns that have only one valence.

**There is no ethical implementation, and here is the proof.** Take the Disclosure Test (frameworks/03): narrate the pattern's DNA chain to the user in plain language. "*We have made cancelling four pages long because we know that some proportion of people who intend to cancel will give up partway through, and those people will pay us for at least one more month.*" Every user who reads that sentence and is still in the flow will complete the flow. The pattern does not survive its own explanation for a

single screen. Compare C-01, where the sentence "*we show a streak because you told us you want to practise daily*" leaves the streak entirely intact.

**The strongest available defence, stated fairly, and why it fails.** The defence is: *cancellation is a consequential, sometimes accidental action; friction protects users from cancelling by mistake; and an offer at the moment of cancellation is legitimate commerce, since the user may genuinely prefer a cheaper plan to leaving.* Every clause of that is true, and each is a real design consideration. The defence fails on **asymmetry**, and only on asymmetry. If accidental cancellation is the concern, then accidental *enrolment* is a strictly larger concern, since enrolment costs money and cancellation saves it. A company that genuinely believed in protective friction would apply it to the enrolment path first. Not one does. The friction is applied exclusively where it protects revenue, and that selectivity is not a side-effect of the design; it *is* the design.

**And note what the Amazon record does to the mens rea question.** The usual defence of deceptive patterns is that no one intended harm, that each screen was a locally-reasonable A/B-tested decision, and that the aggregate was nobody's plan. That defence is available for most of the patterns in this taxonomy and this document takes it seriously (see Behavioral Debt). It is **not** available here. Amazon named the flow after an epic poem about a very long journey. Somebody knew.

#### ETHICAL ALTERNATIVE

The counterpart is **I-03** Frictionless Exit, and it is not a platitude: it is a shipped, documented practice at named companies (see that entry).

1. **Enforce symmetry, and measure it.** Publish the Asymmetry Ratio as an internal metric. Steps to cancel must not exceed steps to subscribe. This is now the *legal* standard imposed on Amazon by the FTC settlement: a clear cancel button, and cancellation through the same medium used to enrol.
2. **One offer, once, skippable.** A retention offer is legitimate commerce. It becomes H-02 in its manipulative sense the moment it is unskippable or repeated. The test: is "no thanks, cancel" always visible on the same screen as the offer?
3. **Cancel means cancelled, immediately, with a receipt.** No re-affirmation loops. Send an email confirming the cancellation and the final billing date. This one change would have made the Amazon case impossible.
4. **Do not make the user hunt for the exit.** Cancellation lives in the same place as billing, which is where the user looks.
5. **Instrument the honest metric.** Report *cancellation-flow abandonment rate* to the board alongside churn. A company that is genuinely retaining users will have a low abandonment rate. A company that is trapping them will have a high one, and will now have to explain it. This costs nothing to build and is the single most uncomfortable number an executive could be asked for, which is why almost no one collects it.

#### REGULATORY STATUS

**This is the most heavily regulated pattern in the taxonomy, and the regulatory picture is genuinely unstable as of July 2026.**

- **United States, enforcement.** *FTC v. Amazon.com, Inc.* settled 25 September 2025: **$2.5bn** ($1bn civil penalty + $1.5bn consumer refunds), under **FTC Act §5(a)** and **ROSCA**. Injunctive relief requires a clear cancellation button and cancellation in the same medium as enrolment. Refunds cover consumers affected between 23 June 2019 and 23 June 2025.
- **United States, rulemaking, and the reversal.** The FTC's revised Negative Option Rule (the "click-to-cancel" rule), which would have required a simple cancellation mechanism for all online subscriptions and separate affirmative consent for negative-option features, was **vacated in its entirety by the U.S. Court of Appeals for the Eighth Circuit on 8 July 2025**, days before it was to take full effect. The vacatur was on **procedural grounds**: the FTC failed to conduct a preliminary regulatory analysis after an ALJ determined the rule's economic impact would exceed $100 million. The court did not hold that the substance was unlawful. The FTC retains its §5 and ROSCA authority (as the Amazon settlement, reached *after* the vacatur, demonstrates), and state automatic-renewal laws remain in force. The FTC has moved to revive the rule.
- **European Union.** The term "dark patterns" was codified into EU law in 2022 through the **Digital Services Act**, the **Digital Markets Act**, and the Data Act proposal. (Source: Gray et al. 2024.) The DSA's specific prohibition on deceptive interface design, article number and wording, was **not retrieved for this document**: CITATION NEEDED .
- **United States, state law.** "Dark patterns" are codified in the California **CPRA**. (Source: Gray et al. 2024.)
- **India.** The Department of Consumer Affairs released draft dark-pattern guidelines in late summer 2023 and finalised them in November 2023. (Source: Gray et al. 2024.)
- **Academic legal analysis.** Luguri & Strahilevitz (2021) argue that many deceptive patterns already violate federal and state unfair-and-deceptive-practices statutes without new legislation, and that consent obtained under them may be voidable under contract-law principles. They propose that deceptive-pattern audits become part of the FTC's consent-decree process. The Amazon settlement is broadly consistent with their thesis.

## RELATED PATTERNS

- H-01 Cancellation Maze, the specific instantiation of this asymmetry in a subscription-cancellation flow. F-09 is the *structure*; H-01 is the *flow*.
- F-10 Forced Continuity ATTR VERIFIED and F-11 Concealed Subscription, the enrolment-side mechanisms that fill the motel.
- F-08 Confirmshaming ATTR VERIFIED and H-03 Guilt-Framed Exit, the copy on the exit path.
- H-02 Exit Offer, H-10 Pause as Trap, H-11 Support Obstruction, H-04 Deletion Obstruction: the components of the maze.
- F-13 Basket Sneaking ATTR VERIFIED , F-12 Preselected Add-on: how the user got in without deciding to.
- I-03 Frictionless Exit and I-08 Reversible Commitment: the ethical counterparts, and the debt repayment.

## SOURCES

Verified project sources (see research/sources/):

- **Brignull, H.**, deceptive.design. Origin of the name (site started 2010; the current published typology has 18 types and includes "Hard to cancel"). https://www.deceptive.design/types
- **Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A.L. (2018).** "The Dark (Patterns) Side of UX Design." CHI '18, Paper 534. DOI 10.1145/3173574.3174108. **Table 1 reproduces Brignull's typology of 11 types, "Roach Motel" among them.** This is the retrieved source for the attribution. The paper's five strategy categories include **Obstruction**, defined verbatim as "Making a process more difficult than it needs to be, with the intent of dissuading certain action(s)."
- **Luguri, J., & Strahilevitz, L.J. (2021).** "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006. (n = 1,963 and n = 3,777; 11.3% / 25.8% / 41.9%; obstruction among the most effective strategies; "Decision architecture, not price, drove consumer purchasing decisions." All quoted figures confirmed verbatim from the article body.)
- **Mathur, A., et al. (2019).** "Dark Patterns at Scale." *PACM HCI* 3(CSCW), Art. 81. DOI 10.1145/3359183. ("Hard to Cancel" is the sole type in their Obstruction category; 1,818 instances across ~11,000 shopping sites.)
- **Thaler, R.H., & Sunstein, C.R. (2021).** *Nudge: The Final Edition*. Yale University Press. ("Sludge." Framework only; no effect size.)
- **Gray, C.M., Santos, C.T., Bielova, N., & Mildner, T. (2024).** "An Ontology of Dark Patterns Knowledge." CHI '24. DOI 10.1145/3613904.3642436. (Source for the DSA / DMA / CPRA / India codification timeline.)

Sources:

- Federal Trade Commission, "FTC Secures Historic $2.5 Billion Settlement Against Amazon," 25 September 2025. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-secures-historic-25-billion-settlement-against-amazon (the FTC page returned HTTP 403 to direct retrieval on ; its contents are relied on here as reported by the law-firm analyses below, which quote it).
- FTC case page, *FTC v. Amazon.com, Inc. (ROSCA)*, matter 2123050. https://www.ftc.gov/legal-library/browse/cases-proceedings/2123050-amazoncom-inc-rosca
- Katten, "FTC's Landmark $2.5 Billion Amazon Settlement Highlights Ongoing Focus on 'Dark Patterns'." https://quickreads.ext.katten.com/post/102l7e6/ftcs-landmark-2-5-billion-amazon-settlement-highlights-ongoing-focus-on-dark-p (source for the "Iliad Flow" four-page / six-click / fifteen-option description and the two-click enrolment).
- Goodwin, "FTC Settles with Amazon Over Enrollment and Cancellation Processes for Prime Memberships," September 2025. https://www.goodwinlaw.com/en/insights/blogs/2025/09/ftc-settles-with-amazon-over-enrollment-and-cancellation-processes-for-prime-memberships (source for the "End Membership" button not ending the membership, and the three re-affirmations).
- FTC, Amazon refunds page. https://www.ftc.gov/enforcement/refunds/amazon-refunds
- Cooley, "Click to Cancel Just Got Cancelled: Eighth Circuit Vacates Entirety of FTC's Negative Option Rule," 11 July 2025. https://www.cooley.com/news/insight/2025/2025-07-11-click-to-cancel-just-got-cancelled-eighth-circuit-vacates-entirety-of-ftcs-negative-option-rule

- WilmerHale, "Eighth Circuit Vacates the FTC's 'Click to Cancel' Rule, but Federal and State Regulators Likely to Remain Active," 1 August 2025. https://www.wilmerhale.com/en/insights/client-alerts/20250801-eighth-circuit-vacates-the-ftcs-click-to-cancel-rule-but-federal-and-state-regulators-likely-to-remain-active

**Note on the trial detail.** The "35 million non-consensual enrolments" figure is attributed in reporting to a statement by FTC counsel at trial, citing Amazon's own estimation. It is reported here as such, and not as an independently verified count.

G-01

# Consent Asymmetry

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| G, Consent, Data and Privacy | Trust | Manipulative | ●●●●● 5/5 |

## DEFINITION

Consent asymmetry is the deliberate imposition of unequal cost on the two answers to a consent question. Accepting is one click. Refusing is several: a second screen, a list of vendors, a set of toggles, a "manage preferences" detour, a "confirm my choices" button at the end of it. Both options exist, so the interface is formally compliant with any rule that requires a choice to be offered. But the *prices* of the two answers differ, and because they differ, the aggregate distribution of answers is set by the designer rather than by the users. The pattern is not the absence of a decline option. It is the **taxation of the decline option**. Identify it in the wild by counting: if accepting takes one action and refusing takes more than one, the interface has an opinion about your privacy and is charging you to disagree with it.

Note the family relationship. G-08 Consent Nudge biases the choice with **visual weight** (a bright accept, a grey decline). G-01 biases it with **cost in clicks**. They co-occur constantly, and G-01 is the more serious of the two, because a visual bias can be overcome by a determined user at no cost, and a click tax cannot.

## PATTERN DNA

**Goal** maximise the rate of consent to tracking, personalisation, or data sharing, on which advertising revenue depends → **Trigger** a legally-mandated consent moment, most often the first page load in a jurisdiction with a consent requirement. Note the irony and hold on to it: **the trigger for this pattern is a law designed to protect the user** → **Interface** a modal, with a single prominent "Accept all" and a "Manage options" or "More choices" link that leads to a second screen of per-purpose toggles, defaulted on, ending in a further confirmation click → **Bias** effort discounting and satisficing: the user is not choosing between privacy and tracking, they are choosing between one click and six, on their way to something else they actually came for → **Emotion** impatience, mild irritation, and above all *indifference*, which is the state the design is cultivating → **Decision** "just make this go away" → **Behavior** Accept all → **Reward** the banner disappears and the user reaches the content they came for. This is a real reward, delivered instantly, and it is the only one on offer → **Business metric** consent rate, which converts directly into addressable ad inventory and CPM.

**The link that breaks on disclosure:** *Interface*. Equal-weight, equal-cost buttons destroy the pattern outright. This is unusually clean: there is no copy change, no education campaign, no dark art. **One symmetric pair of buttons and the pattern is gone.** The fact that a fix this cheap has required regulators to levy nine-figure fines to obtain is the most eloquent available evidence about what the pattern is worth to the companies that use it.

## PSYCHOLOGICAL BASIS

1. **Effort as a price, and the user is not shopping.** The user did not come to the site to make a privacy decision. The consent banner is an obstacle between them and their goal, and their objective in that moment is *to remove the obstacle*, not to optimise their data footprint. Under those conditions, the cheapest path wins, essentially regardless of its content. This is the core mechanism, and it does not require the user to be irrational, misinformed, or biased in any way. A perfectly rational agent, correctly valuing their privacy, will still click Accept if the price of refusing exceeds the marginal value of refusing *this once*. The company only needs that inequality to hold for a few seconds at a time.
2. **Choice architecture and sludge.** Thaler and Sunstein's framing is exactly on point: the decline path is **sludge**, friction that prevents people from doing what they *do* want. Cited as a framework, **not** as an effect size; no pooled nudge effect is asserted anywhere in this document, because the pooled effect is disputed under publication-bias correction.
3. **Obstruction, measured.** Luguri & Strahilevitz (2021) tested deceptive-pattern strategies experimentally on representative American samples (Study 1 n = 1,963; Study 2 n = 3,777). Acceptance of a dubious service rose from **11.3%** (control) to **25.8%** (mild deceptive patterns) to **41.9%** (aggressive), and Study 2 found **obstruction** among the three strategies "particularly likely to manipulate consumers successfully." Their most consequential finding for this entry is that under deceptive patterns **price became immaterial**: "Decision architecture, not price, drove consumer purchasing decisions." Transpose that to consent and the implication is severe: **if architecture can override price, it can certainly override a stated privacy preference**, which is a far weaker commitment than a monetary one. The gap between what users say they want and what they click is routinely read as hypocrisy. Luguri & Strahilevitz give a better explanation: it is architecture.
4. **Consent fatigue.** Repeated across every site, the effect compounds into indifference. See G-11.

#### INTERFACE EXPRESSION

- A single, saturated, high-contrast **"Accept all"** button.
- A secondary path labelled to sound like *more work* rather than like a *choice*: "Manage options," "Customise," "More choices," "Purposes." The label is doing work: it describes a task, not an answer.
- The second screen presents per-purpose toggles, frequently **defaulted to on** (which is G-02 Preselected Opt-in, a separate pattern stacked on this one).
- A vendor list, sometimes running to hundreds of entries, each with its own toggle (G-11 Consent Fatigue).
- A terminal "Save and continue" that is *itself* an additional click, so that even a user who reaches the end and toggles everything off pays a higher price than the user who accepted at step one.
- **The measurement:** one click to accept, four to seven to refuse. That ratio is the pattern.

#### REAL-WORLD EXAMPLES

**CNIL v. Google and Facebook, January 2022.** This is the definitive documented case, and it is documented precisely because the regulator counted the clicks.

- On **6 January 2022** France's data-protection authority, the CNIL, fined **Google €150 million** (comprising €90m against Google LLC and €60m against Google Ireland Ltd) and **Facebook €60 million**.

- The finding was not that consent was absent. It was that consent was **asymmetric**. Per CNIL's own account of the decisions, google.fr and youtube.com offered a button allowing cookies to be accepted immediately, but provided no equivalent means of refusing them: **several clicks were necessary to refuse all cookies, against a single click to accept them.** The same finding was made against facebook.com.
- The legal basis was **Article 82 of the French Data Protection Act** (implementing the ePrivacy Directive), not the GDPR. This is a jurisdictional detail with real consequence: it allowed CNIL to act nationally rather than through the GDPR's one-stop-shop mechanism.
- CNIL ordered both companies to provide French users with a means of refusing cookies **"as simple as the existing means of accepting them,"** within **three months**, on pain of **€100,000 per day** of delay.
- Google complied by adding a refusal button, titled "Only allow essential cookies," adjacent to the acceptance button. CNIL closed the injunction proceeding on **13 July 2023**.

Read the shape of that story carefully, because it is the whole entry in miniature. The remedy was **one button**. It took a €150 million fine and a daily penalty to obtain one button. Whatever the asymmetry was worth, it was worth more than the button cost to build.

**A note on prevalence.** Consent asymmetry is not confined to two companies; it is the default condition of the consent-banner web. This document does not, however, print a prevalence figure, because no source establishing one was retrieved. CITATION NEEDED for any measurement of how common asymmetric consent interfaces are. Mathur et al. (2019) measured deceptive patterns across ~11,000 shopping sites but their scope was e-commerce dark patterns, not cookie consent, and their categories do not include this one.

No current cookie banner is described from memory, per the evidence rule in PROJECT.md §7. Banners change, frequently in direct response to enforcement of the kind described above.

#### BUSINESS RATIONALE

Consent rate is the input to addressable advertising inventory, and addressable inventory is worth a large multiple of non-addressable inventory. A consent banner is therefore not a compliance artifact; it is a **revenue-critical funnel step**, and it is optimised exactly as a checkout funnel is optimised, by the same teams, with the same tools, against the same kind of A/B test.

That framing explains the pattern's persistence better than any theory of malice. Once a legal requirement to obtain consent exists, and once consent is a funnel, the funnel will be optimised. The company does not need to decide to manipulate anyone. It needs only to run the experiment, observe that the one-click variant converts better, and ship it. Luguri & Strahilevitz make precisely this point about the industry as a whole: deceptive patterns are "presumably proliferating because firms' proprietary A-B testing has revealed them to be profit maximizing."

There is a second-order lesson here that this document wants to state plainly. **A consent requirement without a symmetry requirement is not a protection; it is a specification for a dark pattern.** The law created the banner. The law did not, initially, say the two answers must cost the same. Everything that followed was predictable.

### BEHAVIORAL COST INDEX

Scored per frameworks/02. Scores reproduced from the worked example in that framework document.

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 2 | The extra clicks are real but small. The user is aware of them. This is the axis on which the pattern looks trivial, and it is why the pattern is so often dismissed. |
| Attention | 3 | It requires vigilance the user did not budget for: they must notice that "Manage options" is the refusal path, and that the toggles behind it are pre-set. |
| Memory | 1 | Little to hold in mind. |
| Trust | 4 | The asymmetry is perfectly legible once it is seen, and once seen it reads as contempt. The user would feel betrayed on realising the interface was engineered to tax their disagreement, and that realisation changes their behavior toward the brand. |
| Money | 0 | None. |
| **Privacy** | **5** | **The load-bearing score. The design's entire function is to obtain consent that would not be given under symmetric conditions.** The user is unaware of the gap between what they clicked and what they would have clicked at equal cost; awareness would change their behavior; and the design works *precisely because* of that unawareness. If users did the arithmetic, the pattern would yield nothing. |
| Emotion | 2 | Irritation. |

**Max = 5 → Manipulative.** Not because we disapprove of it, but because it fails the test: it works *because* the user does not do the arithmetic.

Observe the shape of this profile and compare it with F-09. F-09 is a **Money 5 / Trust 5** pattern with a Privacy 0. G-01 is a **Privacy 5** pattern with a Money 0. Both are manipulative, and both would be assigned the same "severity: 5" by any conventional taxonomy, which would tell you nothing. The BCI tells you they are entirely different harms, which is the point of having seven axes instead of one.

### VALENCE ANALYSIS

**There is no ethical implementation, and the reason is definitional.** If accept and decline cost the same, the pattern is not G-01. It is **I-02** Symmetric Consent, which is a different entry with a different ID. G-01 *is* the asymmetry; remove the asymmetry and you have removed the pattern. This is why TAXONOMY.md gives it a valence range with a single value.

**The defence, stated at its strongest.** It runs like this: *the choices genuinely are not symmetric. "Accept all" is one decision; "manage preferences" is many. A user who wants to accept some purposes*

*and not others needs a second screen, because that granularity cannot fit on the first. The extra clicks reflect genuine complexity, not obstruction. Moreover, users have repeatedly demonstrated by their behavior that they do not care much about this, and forcing them through a granular privacy dialog serves an abstraction ("privacy") over their revealed preference ("let me read the article").*

**Why the defence fails, and it fails cleanly.** The defence would justify a *third* option, not the removal of the second. Granularity is a real user need, and nothing about a symmetric interface prevents serving it. The honest layout is **three affordances**: Accept all (one click), **Reject all (one click)**, Manage preferences (many clicks, for the minority who want granularity). CNIL's remedy was exactly this, and Google implemented exactly this ("Only allow essential cookies," adjacent to the accept button), and the sky did not fall. The complexity argument is a defence of the *third* button being expensive. It has never been a defence of the *second* button being absent, and the second button is the entire allegation.

The second half of the defence, the revealed-preference argument, is the more interesting one and it deserves a real answer rather than a dismissal. It says: users click accept, therefore users consent, therefore where is the harm? The answer is Luguri & Strahilevitz. Their experiments show that **decision architecture overrides price**, and price is a far stronger preference signal than a privacy attitude. If architecture can make people buy something they did not want, at a price they did not care about, then a click made under an asymmetric architecture is simply **not evidence of a preference**. It is evidence of an architecture. The revealed-preference defence assumes the very thing the experiment refutes: that the click reveals a preference at all.

**Under the Disclosure Test** (frameworks/03), the chain snaps at *Interface*: tell the user "we made refusing six clicks because at one click, most people refuse," and the click they were about to make becomes a different click. That is the structural definition of manipulation used throughout this document.

#### ETHICAL ALTERNATIVE

**I-02** Symmetric Consent. Concretely, and none of this is hypothetical, since it is what CNIL's injunction obtained:

1. **Reject all, one click, on the first screen, adjacent to Accept all.** Equal size, equal contrast, equal position in the reading order. No second screen, no confirmation, no "are you sure."
2. **Manage preferences as a third option**, for the minority who want granularity. It may be as long as it needs to be, because nobody is being taxed into it.
3. **Toggles default to off** (**I-01** Honest Default). A default is a decision the company made on the user's behalf; the honest default is the one a reasonable user would choose if they were paying full attention. This closes the G-02 stacking.
4. **Do not re-ask a user who declined.** Re-asking is **G-05** Permission Nagging, and it converts a one-time tax into a recurring one.
5. **Publish the consent rate before and after.** This is the test of good faith, and it is a hard one. A symmetric banner will produce a materially lower consent rate than an asymmetric one. That difference is exactly the quantity of consent the old design was extracting rather than obtaining, and a company willing to publish it is a company that has genuinely repaid the debt (frameworks/03: any ethical-design initiative that costs the company nothing is not repaying anything).

## REGULATORY STATUS

**This is one of the few patterns with direct, monetised enforcement precedent.**

- **France (CNIL), 6 January 2022.** Google fined **€150m** (€90m Google LLC, €60m Google Ireland), Facebook fined **€60m**, under **Article 82 of the French Data Protection Act** (implementing the ePrivacy Directive, not the GDPR). Finding: several clicks were required to refuse cookies against one to accept. Injunction: provide a refusal mechanism "as simple as the existing means of accepting," within three months, subject to **€100,000 per day** thereafter. Injunction closed against Google on 13 July 2023 after compliance.
- **European Union.** The term "dark patterns" was codified into EU law in 2022 via the **Digital Services Act**, the **Digital Markets Act**, and the Data Act proposal. (Source: Gray et al. 2024.) The DSA is understood to contain a prohibition on deceptive interface design by online platform providers, but **the specific article number and its operative wording were not retrieved for this document** and are therefore CITATION NEEDED. The statutory text must be pulled from EUR-Lex before publication; it is not paraphrased here from memory.
- **GDPR.** Consent must be freely given, specific, informed and unambiguous. An asymmetric interface is a direct challenge to "freely given." CITATION NEEDED for the specific article text and for European Data Protection Board guidance on cookie banners; neither was retrieved for this document.
- **United States.** "Dark patterns" are codified in the California **CPRA**. (Source: Gray et al. 2024.) There is no federal analogue to the CNIL action.
- **India.** Department of Consumer Affairs dark-pattern guidelines, drafted late summer 2023 and finalised November 2023. (Source: Gray et al. 2024.)
- **Academic legal analysis.** Luguri & Strahilevitz (2021) argue many deceptive patterns already violate existing federal and state unfair-and-deceptive-practices law, and that consent given under them may be voidable under contract-law principles. On the facts of the CNIL decisions, that argument looks conservative rather than bold.

## RELATED PATTERNS

- I-02 Symmetric Consent, the ethical counterpart, and definitionally the negation of this pattern.
- G-02 Preselected Opt-in and G-03 Default Over-disclosure ATTR VERIFIED, which are almost always stacked behind the "Manage options" screen.
- G-08 Consent Nudge, the visual-weight sibling. G-01 taxes the decline; G-08 merely hides it.
- G-11 Consent Fatigue, the aggregate effect across every site the user visits, and the reason the pattern's per-site cost understates its true cost.
- G-05 Permission Nagging ATTR VERIFIED, repetition of a rejected request.
- G-12 Ambiguous Wording, G-04 Bundled Consent: adjacent consent-quality failures.
- F-09 Roach Motel, the same asymmetry logic applied to exit rather than to consent. The two patterns share a structure and differ only in what is being taxed.
- I-01 Honest Default.

## SOURCES

Verified project sources (see research/sources/):

- **Luguri, J., & Strahilevitz, L.J. (2021).** "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006. (Study 1 n = 1,963; Study 2 n = 3,777; control 11.3%, mild 25.8%, aggressive 41.9%; obstruction among the most effective strategies; "Decision architecture, not price, drove consumer purchasing decisions"; deceptive patterns proliferate because firms' A-B testing shows them to be profit-maximising. All figures confirmed verbatim from the article body.)
- **Thaler, R.H., & Sunstein, C.R. (2021).** *Nudge: The Final Edition*. New Haven: Yale University Press. ("Choice architecture," which the publisher credits the authors with inventing; and "sludge." Cited as a framework. **No nudge effect size is asserted**, because the pooled effect is disputed under publication-bias correction.)
- **Gray, C.M., Santos, C.T., Bielova, N., & Mildner, T. (2024).** "An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building." CHI '24. DOI 10.1145/3613904.3642436. (Source for the DSA / DMA / Data Act / CPRA / India codification timeline. Cite the Crossref version of record, not the February 2024 preprint, whose title and author order differ.)
- **Gray, C.M., et al. (2018).** "The Dark (Patterns) Side of UX Design." CHI '18. DOI 10.1145/3173574.3174108. (Interface Interference: "Manipulation of the user interface that privileges certain actions over others," quoted verbatim. This is the strategy category G-01 belongs to.)
- **Brignull, H.**, deceptive.design /types. (Current typology of 18 includes "Visual interference" and "Trick wording," the neighbours of this pattern.)

Sources:

- TechCrunch, "France spanks Google $170M, Facebook $68M over cookie consent dark patterns," 6 January 2022. https://techcrunch.com/2022/01/06/cnil-facebook-google-cookie-consent-eprivacy-breaches/ (fine amounts; the CNIL quotation "the methods of collecting consent proposed to users, as well as the lack of clarity of information provided to them, constitute violations of Article 82 of the French Data Protection Act"; the three-month injunction and the €100,000-per-day penalty).
- CNIL (French data protection authority), decision **SAN-2021-023** (Google) and the parallel Facebook decision, 31 December 2021, announced 6 January 2022. CNIL's own English announcement page (cnil.fr/en/cookies-cnil-fines-google-total-150-million-euros-and-facebook-60-million-euros-non-compliance) returned **HTTP 404** on and the closure-of-injunction page returned **HTTP 403**; the decision's content is therefore relied on here as reported by the secondary sources cited, and the primary deliberation should be re-retrieved from CNIL's Légifrance-hosted text before final publication. **Flagged as an outstanding verification task.**
- CNBC, "Google hit with 150 million euro French fine for cookie breaches," 6 January 2022. https://www.cnbc.com/2022/01/06/google-hit-with-150-million-euro-french-fine-for-cookie-breaches.html
- Hunton Andrews Kurth, "CNIL Fines Big Tech Companies 210 Million Euros for Cookie Violations." https://www.hunton.com/privacy-and-cybersecurity-law-blog/cnil-fines-big-tech-companies-210-million-euros-for-cookie-violations (split of the Google fine between Google LLC and Google Ireland; the "Only allow essential cookies" remedy and the July 2023 closure).

H-01

# Cancellation Maze

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| H, Retention and Exit Friction | Exit | Manipulative | ●●●●● 5/5 |

## DEFINITION

A cancellation maze is a multi-step flow interposed between a user's decision to cancel a subscription and the cancellation actually taking effect, composed of screens each of which is individually defensible and which are collectively designed to exceed the user's supply of patience. Its components are re-confirmations, retention offers, benefit reminders, pause and downgrade diversions, exit surveys, channel changes, and, at the limit, a requirement to speak to a human. The definitional property is not length as such. It is that **the flow is longer than it needs to be to establish the user's intent, and the surplus length is where the revenue is.** The user's intent is established at the first click; everything after it is an attempt to spend that intent down.

The distinction from F-09 Roach Motel is worth stating precisely, because the two are often collapsed. **F-09 is the structural asymmetry** between the entry path and the exit path, and it is diagnosed by a ratio. **H-01 is the flow itself**: the specific sequence of screens, and the specific psychological work each one does. F-09 tells you the company is opposed to your leaving. H-01 tells you how.

## PATTERN DNA

**Goal** convert a cancellation into a retention, or failing that, into a delay, which is worth at least one more billing cycle → **Trigger** the user's own decision to leave. This pattern is unique, with F-09, in being triggered entirely by the user, which is why it is the only place in the product where the company's interest and the user's are in *open* conflict rather than latent conflict → **Interface** a sequence of screens, each presenting the user with something other than the cancel button they are looking for → **Bias** several, stacked deliberately: sunk cost in the flow already traversed; present bias, since abandoning is cheaper right now than finishing; loss framing on the benefits screen ("you will lose your history, your discounts, your saved items"); and depletion, which is the real mechanism and is simply the exhaustion of a finite motivational resource → **Emotion** irritation, then fatigue, then shame if the copy is confirmshaming (F-08, H-03), then resignation → **Decision** "I will deal with this later," which is the design's target output. Note that the design does **not** need the user to decide to stay. It needs only that they fail to finish leaving → **Behavior** flow abandonment → **Reward** none. **There is no user-side reward at any link of this chain**, which is diagnostic → **Business metric** churn rate, retained MRR.

**The link that breaks on disclosure:** *Emotion*, and specifically the fatigue. Show the user a progress indicator at the top of the flow ("step 1 of 4, you may leave at any time and your cancellation will still be processed") and the flow's power evaporates, because the user can now price the remaining cost and will simply pay it. **The maze depends on the user not knowing how much further there is to go.** That is why cancellation flows, uniquely among multi-step flows in consumer software, almost never carry the progress indicator that every checkout flow carries as a matter of course. The omission is not an oversight.

### PSYCHOLOGICAL BASIS

1. **Obstruction, and it is experimentally established.** Luguri & Strahilevitz (2021) ran two randomized experiments on representative samples of American consumers (Study 1 n = 1,963; Study 2 n = 3,777). Acceptance of a dubious service rose from **11.3%** in the control to **25.8%** under mild deceptive patterns and **41.9%** under aggressive ones (the paper renders these as 228% and 371% increases). Study 2 found that **hidden information, trick questions, and obstruction strategies were "particularly likely to manipulate consumers successfully."** Obstruction is the mechanism of H-01, and this is as direct an empirical warrant as any pattern in this taxonomy has.

The authors' further finding bears repeating in this context because it demolishes the standard defence: under deceptive patterns, **"Decision architecture, not price, drove consumer purchasing decisions."** If architecture can override *price*, the claim that a user who really wanted to cancel would have persisted is not a claim about users. It is a claim about architecture, and it is false.

1. **Sludge.** Thaler and Sunstein's own term for friction that keeps people from what they *do* want. A cancellation maze is the purest possible instance of it. Cited as a framework, not as an effect size.
2. **Obstruction as a designer strategy.** Gray et al. (2018) define **Obstruction**, verbatim, as "*Making a process more difficult than it needs to be, with the intent of dissuading certain action(s).*" Note that their definition contains an intent clause. This is one of the very few patterns for which intent is not an inference: a four-page cancellation flow does not arise from local optimisation, it arises from someone specifying four pages.
3. **Depletion.** The maze works by outlasting motivation. The general claim that self-control or persistence is a depletable resource is **contested in the psychological literature** and its replication status is poor. CITATION NEEDED, and the claim is not relied on here. Fortunately it does not need to be: the Luguri & Strahilevitz obstruction result stands on its own without any theory of *why* obstruction works.

### INTERFACE EXPRESSION

The canonical sequence, in the order it is usually encountered:

1. **The hidden entrance.** Cancellation is not where billing is. It is under Help, or Account, or a link at the bottom of a settings page.
2. **The reason survey.** "Why are you leaving?" presented before the cancellation, not after. An exit survey after cancellation is research. An exit survey *before* it is a gate.
3. **The benefits reminder.** A list of everything the user will lose (D-07, H-06).
4. **The offer** (H-02). A discount, a free month. Legitimate commerce, if skippable on the same screen. A delay tactic if not.
5. **The pause** (H-10). Framed as helpful. Keeps the account, and the billing relationship, alive.
6. **The downgrade** (H-07). A cheaper tier, offered because a smaller payment beats no payment.
7. **The confirmshaming** (F-08, H-03). "No thanks, I don't want to save money."
8. **The re-affirmation loop.** The user clicks a button that ends the membership. The membership does not end. They are asked again. And again.

9. **The channel change.** Cancel by phone. During business hours. In a specific timezone (H-11).

**The single most diagnostic feature of the whole family: a button labelled to end the relationship that does not end the relationship.** Everything else is a matter of degree. That one is a lie in the interface.

#### REAL-WORLD EXAMPLES

**Amazon Prime, FTC v. Amazon.com, Inc. (ROSCA).** The best-documented cancellation maze in existence, because it went to trial and the internal artifacts became public.

- Amazon's internal name for the Prime cancellation flow was the **"Iliad Flow"**, a reference to a long and arduous epic. Trial reporting describes it as a **four-page, six-click, fifteen-option** sequence, with repeated diversions including discount offers and reminders of Prime benefits. **Enrolment took two clicks.**
- The court found a dispute of material fact as to whether Amazon offered a simple cancellation mechanism, because clicking the **"End Membership"** button **did not end the membership**: customers had to reaffirm the desire to cancel **three times** before the cancellation became effective. That is component 8 above, documented in a federal proceeding.
- Per the FTC's counsel at trial, by Amazon's own estimation there were more than **35 million non-consensual Prime enrolments** over a seven-year period. (Reported as counsel's characterisation of Amazon's own estimate; not independently verified here.)
- On **25 September 2025**, on the fourth day of a jury trial that began on 22 September 2025, Amazon settled for **$2.5 billion**: a **$1 billion civil penalty** and **$1.5 billion in consumer refunds**, under **FTC Act §5(a)** and **ROSCA**. Refunds cover consumers affected between 23 June 2019 and 23 June 2025.
- The settlement's injunctive relief requires Amazon to present a **clear cancellation button** and to provide a **simple cancellation mechanism through the same medium the customer used to enrol.** That remedy is, precisely, **I-03** Frictionless Exit, imposed by court order.

**Regulatory naming of the pattern.** Brignull's current published typology names **"Hard to cancel"** as one of its 18 types. Mathur et al. (2019) found **"Hard to Cancel"** to be the sole type populating their **Obstruction** category, in a crawl of ~11,000 shopping websites that identified 1,818 dark-pattern instances across 15 types and 7 categories. Gray et al. (2024) harmonised ten taxonomies into 64 types across three levels, and cite the **$245 million FTC judgment against Epic Games (Fortnite)** as an enforcement exemplar.

No other product's cancellation flow is described here. Cancellation flows are altered frequently, often in direct response to the enforcement described above, and a remembered flow is not evidence about a current one (PROJECT.md §7).

#### BUSINESS RATIONALE

The maze is the highest-ROI feature in a subscription business, and it is worth being exact about why, because the ethics argument is not persuasive to anyone who has not first understood the finance.

Let *a* be the fraction of users who begin the cancellation flow and abandon it. Every one of those users pays at least one more billing cycle, and in practice many pay indefinitely, because a user who abandoned once has demonstrated that their intent-to-cancel is below the threshold the flow imposes, and the flow does not get shorter. The cost of building the maze is one engineering sprint. The revenue is *a* × (subscribers attempting cancellation) × (price) × (cycles), recurring. There is no other feature with that ratio.

And the pattern is **self-validating on the dashboard**. Churn falls after the flow ships. Retention rises. Every standard analytics stack reports this as a product win, and **not one of them distinguishes "users chose to stay" from "users failed to leave."** The variable that would separate the two, cancellation-flow abandonment rate, is not part of any standard funnel report and is almost never asked for. This is Behavioral Debt (frameworks/03) in its most literal form: the instrumentation books the interest payment and reports it as income.

Which is exactly why the Amazon settlement matters as an economic fact rather than a moral one. **$2.5 billion is the first number in this pattern's history that appears on the same balance sheet as the revenue it retained.** Until an enforcement expectation is priced in, the maze is a rational build. That is not a cynical observation, it is the decision structure, and any argument against the pattern that does not engage with it will be ignored by the people who authorise it.

#### BEHAVIORAL COST INDEX

Scored per frameworks/02. **One column only.** A cancellation maze whose costs were fully visible to the user would not function as a cancellation maze, because the user would simply price the remaining screens and pay them. The absence of an ethical column is not laziness; it is the result.

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 4 | Multiple screens, sometimes multiple sessions, sometimes a phone call in business hours. The user did not budget for it, would not have accepted it if told in advance, and the total is deliberately not disclosed at the outset (see: no progress indicator). |
| Attention | 4 | The flow is adversarial. Every screen presents something other than the thing the user came for, and the user must remain vigilant not to be diverted into a pause, a downgrade, or an offer they did not want. Attention here is *taken*, not given. |
| Memory | 3 | The user must remember that they may not actually have cancelled, and must remember to verify. This axis is doing real work here and is close to zero in most other patterns. |
| Trust | 5 | The betrayal is complete and unambiguous. The user learns that the company built a machine specifically to defeat them, and that the button labelled "End Membership" was not telling the truth. There is no recovering from this, and users do not forget it. |
| **Money** | **5** | **Load-bearing.** The user is charged for a service they decided to stop, after taking an action they believed had stopped it. They are unaware they are still being billed; awareness would immediately change their behavior; and the design exists precisely in order to produce that unawareness. The FTC's Amazon refund pool, $1.5 billion, is a market estimate of this axis. |
| Privacy | 0 | The pattern takes time and money. It does not take data. |
| Emotion | 4 | Frustration, humiliation, and (where confirmshaming is present, F-08/H-03) shame, all induced on purpose. The user is only partly aware that the affect is engineered rather than incidental. |

**Max = 5 → Manipulative.** Five of seven axes score 3 or above, which is unusual: most manipulative patterns are narrow, spiking on one or two axes. **The cancellation maze is broad.** It is one of the most expensive patterns in the taxonomy on total Behavioral Cost Load, and its Opacity (count of axes at 4 or 5) is **4**, which is the highest of the nine entries written so far.

### VALENCE ANALYSIS

**There is no ethical implementation.** Not because we disapprove, but because of what happens when you remove the harm: a cancellation flow that is as short as the intent requires, with a skippable offer, is

not a "gentle H-01." It is **I-03**, a different pattern with a different ID. The maze *is* the surplus length. Delete the surplus and the pattern is gone.

**The strongest defence, and it must be given fairly, because it contains two true premises.**

*Premise 1: cancellation is consequential and sometimes accidental, so confirmation protects the user.* True. A user who cancels by mistake loses their history, their data, their price grandfathering.

*Premise 2: an offer at the moment of cancellation is legitimate commerce.* Also true. Many cancelling users would genuinely prefer a cheaper plan or a pause to leaving altogether, and a company that never offered one would be failing them.

**The defence collapses on two counts, and it is worth naming both, because only one of them is usually noticed.**

*First, the asymmetry.* If accidental action is the concern, accidental **enrolment** is the graver concern, because enrolling costs the user money and cancelling saves it. A company that genuinely believed in protective friction would install it on the enrolment path first. Amazon enrolled in two clicks and cancelled in six across four pages. Protective friction is applied exclusively where it protects revenue. That selectivity is not an artifact of the design; it is the design.

*Second, and this is the more precise cut: intent is established at the first click, and everything after it is not confirmation but negotiation.* One confirmation establishes intent. Two is redundant. **Three, which is what Amazon required, is not a confirmation mechanism at all**, because no amount of additional confirmation makes the intent clearer than it was at confirmation two. The only function of the third is to give the flow one more chance to lose the user. That is not a philosophical claim. It is arithmetic, and the FTC put a $2.5 billion price on it.

**Under the Disclosure Test** (frameworks/03): "*We made this four pages long because a measurable fraction of people who intend to cancel will give up partway through, and every one of them pays us for at least one more month.*" No user reads that sentence and abandons the flow. The pattern does not survive its own explanation for one screen. Contrast C-01, where the equivalent sentence leaves the pattern fully intact. That contrast is the taxonomy's central claim, and H-01 is one of the two poles it is measured against.

#### ETHICAL ALTERNATIVE

**I-03** Frictionless Exit. This is not aspirational; it is now the **court-ordered standard for Amazon**, and it is documented published policy at 37signals (see the I-03 entry). Concretely:

1. **Cancellation lives where billing lives**, and is reachable in the same number of clicks as subscribing. Publish the Asymmetry Ratio (frameworks/03) internally as a tracked metric.
2. **One confirmation, and it is real.** The button that says the subscription is cancelled cancels the subscription. Send an email receipt with the effective date. This single change would have made the entire Amazon case impossible.
3. **One offer, once, skippable on the same screen.** "Cancel anyway" must be visible at the same moment as the offer, in equal weight (**I-02**). An offer the user can decline in one click is commerce. An offer they must traverse is a toll.

4. **Exit survey after, not before**, and optional. If the goal is genuinely to learn why the user left, ask them once they have left, when their answer is honest and unpressured.
5. **Cancel in the same medium as sign-up.** If they subscribed in an app, they can cancel in the app. This is the FTC's own remedy.
6. **Instrument the honest metric: cancellation-flow abandonment rate**, reported to the board next to churn. A company genuinely retaining users will show a low one. A company trapping them will show a high one, and will have to explain it. It costs nothing to build, which is precisely why nobody builds it.
7. **Pro-rate the unused period, or say plainly that you do not.** (37signals, whose cancellation policy is quoted in **I-03**, does the latter: they state that they do not automatically pro-rate, and invite the user to ask. Honesty about a less generous policy is worth more than silence about a generous-sounding one.)

The test of all seven: **does it cost the company a metric it could have kept?** Every one of them does. That is what distinguishes repayment from marketing.

#### REGULATORY STATUS

**The most heavily enforced pattern in the taxonomy, and the legal position is genuinely in flux as of July 2026.**

- **United States, enforcement.** *FTC v. Amazon.com, Inc.* (ROSCA), settled **25 September 2025** for **$2.5 billion** ($1bn civil penalty + $1.5bn refunds) under **FTC Act §5(a)** and the **Restore Online Shoppers' Confidence Act**. Injunctive relief: a clear cancel button, and cancellation through the same medium as enrolment. Settled on day four of a jury trial.
- **United States, rulemaking, and its reversal.** The FTC's revised **Negative Option Rule** ("click-to-cancel"), which would have mandated a simple cancellation mechanism for all online subscriptions and separate affirmative consent for negative-option features, was **vacated in its entirety by the Eighth Circuit on 8 July 2025**, days before full effect. The vacatur was **procedural**: the FTC failed to conduct a preliminary regulatory analysis after an ALJ found the rule's economic impact would exceed $100 million. The court did not hold the substance unlawful. The FTC has moved to revive it. **Crucially, the Amazon settlement was reached after the vacatur**, which demonstrates that the FTC's §5 and ROSCA authority is undiminished. Companies also remain subject to a patchwork of **state automatic-renewal laws** whose requirements substantially overlap the vacated rule.
- **European Union.** "Dark patterns" codified in 2022 via the **Digital Services Act**, the **Digital Markets Act**, and the Data Act proposal. (Source: Gray et al. 2024.) The DSA's specific provision on deceptive interface design, its article number and its wording, was **not retrieved for this document**: CITATION NEEDED, to be pulled from EUR-Lex before publication rather than paraphrased.
- **United States, state law.** Codified in the California **CPRA**. (Source: Gray et al. 2024.)
- **India.** Department of Consumer Affairs dark-pattern guidelines, drafted late summer 2023, finalised November 2023. (Source: Gray et al. 2024.)
- **Prior art in the literature.** Brignull's published typology names "Hard to cancel" among 18 types. Mathur et al. (2019) measured it as the sole type in their Obstruction category. Luguri & Strahilevitz

(2021) propose that deceptive-pattern audits be built into the FTC's consent-decree process, and argue that consent obtained under deceptive patterns may be voidable under contract-law principles.

### RELATED PATTERNS

- **F-09** Roach Motel ATTR VERIFIED, the structural asymmetry of which this flow is the instantiation. **Read the two entries together.**
- **H-02** Exit Offer, **H-10** Pause as Trap, **H-07** Downgrade Penalty: the diversions inside the maze.
- **H-03** Guilt-Framed Exit and **F-08** Confirmshaming ATTR VERIFIED: the copy.
- **H-11** Support Obstruction: the channel change to phone-only cancellation.
- **H-04** Deletion Obstruction, **H-05** Reactivation Window, **H-06** Data Hostage: the account-level equivalents, applied after the subscription ends.
- **F-10** Forced Continuity ATTR VERIFIED, **F-11** Concealed Subscription: how the billing relationship was established.
- **I-03** Frictionless Exit and **I-08** Reversible Commitment: the ethical counterparts and the debt repayment.

### SOURCES

Verified project sources (see research/sources/):


- **Luguri, J., & Strahilevitz, L.J. (2021).** "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006. (Study 1 n = 1,963; Study 2 n = 3,777; 11.3% / 25.8% / 41.9%, rendered by the paper as 228% and 371% increases; **obstruction** among the three most effective strategies; "Decision architecture, not price, drove consumer purchasing decisions." All figures confirmed verbatim from the article body.)
- **Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A.L. (2018).** "The Dark (Patterns) Side of UX Design." CHI '18, Paper 534. DOI 10.1145/3173574.3174108. (**Obstruction**, verbatim: "Making a process more difficult than it needs to be, with the intent of dissuading certain action(s).")
- **Mathur, A., et al. (2019).** "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites." *PACM HCI* 3(CSCW), Art. 81. DOI 10.1145/3359183. (~11,000 sites; ~53,000 product pages; 1,818 instances; 15 types across 7 categories; **"Hard to Cancel" is the sole type in the Obstruction category**.)
- **Brignull, H.**, deceptive.design /types. ("Hard to cancel," one of 18 published types.)
- **Gray, C.M., Santos, C.T., Bielova, N., & Mildner, T. (2024).** "An Ontology of Dark Patterns Knowledge." CHI '24. DOI 10.1145/3613904.3642436. (10 taxonomies harmonised into 64 types, 3 levels; source for the DSA/DMA/CPRA/India timeline and the $245m FTC judgment against Epic Games.)
- **Thaler, R.H., & Sunstein, C.R. (2021).** *Nudge: The Final Edition*. Yale University Press. ("Sludge." Framework only; no effect size asserted.)


Sources:

- Federal Trade Commission, "FTC Secures Historic $2.5 Billion Settlement Against Amazon," 25 September 2025. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-secures-historic-25-billion-settlement-against-amazon (the FTC page itself returned **HTTP 403** to direct retrieval on ; its contents are relied on as quoted in the law-firm analyses below).
- FTC case page, *FTC v. Amazon.com, Inc. (ROSCA)*, matter 2123050. https://www.ftc.gov/legal-library/browse/cases-proceedings/2123050-amazoncom-inc-rosca
- Katten, "FTC's Landmark $2.5 Billion Amazon Settlement Highlights Ongoing Focus on 'Dark Patterns'." https://quickreads.ext.katten.com/post/102l7e6/ftcs-landmark-2-5-billion-amazon-settlement-highlights-ongoing-focus-on-dark-p (source for the "Iliad Flow" four-page / six-click / fifteen-option description and the two-click enrolment; the 35 million figure attributed to FTC counsel at trial).
- Goodwin, "FTC Settles with Amazon Over Enrollment and Cancellation Processes for Prime Memberships," September 2025. https://www.goodwinlaw.com/en/insights/blogs/2025/09/ftc-settles-with-amazon-over-enrollment-and-cancellation-processes-for-prime-memberships (source for the "End Membership" button not ending the membership and the three required re-affirmations; the same-medium cancellation requirement).
- Cooley, "Click to Cancel Just Got Cancelled: Eighth Circuit Vacates Entirety of FTC's Negative Option Rule," 11 July 2025. https://www.cooley.com/news/insight/2025/2025-07-11-click-to-cancel-just-got-cancelled-eighth-circuit-vacates-entirety-of-ftcs-negative-option-rule
- Latham & Watkins, "Eighth Circuit Vacates FTC's Click-to-Cancel Rule Days Before Compliance Deadline." https://www.lw.com/en/insights/eighth-circuit-vacates-ftc-click-to-cancel-rule-days-before-compliance-deadline (procedural basis: absent preliminary regulatory analysis; ALJ finding of >$100m impact).
- WilmerHale, "Eighth Circuit Vacates the FTC's 'Click to Cancel' Rule, but Federal and State Regulators Likely to Remain Active," 1 August 2025. https://www.wilmerhale.com/en/insights/client-alerts/20250801-eighth-circuit-vacates-the-ftcs-click-to-cancel-rule-but-federal-and-state-regulators-likely-to-remain-active (state automatic-renewal laws survive).

I-03

# Frictionless Exit

| FAMILY | STAGE | VALENCE | SEVERITY CEILING |
|---|---|---|---|
| I, Trust and Autonomy Preservation | Exit | **Ethical** | ●●●●● 0/5 |

## DEFINITION

Frictionless exit is the deliberate construction of a cancellation path that is no harder than the signup path: in the same place, in the same medium, in the same order of magnitude of steps, with no interposed offers the user must traverse, no re-affirmation loops, and no channel change. It is the negation of **F-09** Roach Motel and of **H-01** Cancellation Maze, and it is a **design commitment**, not an absence of design: leaving a cancellation flow unbuilt produces obstruction by default, because the friction accumulates on its own. Someone has to decide that the exit is a first-class path and then defend that decision against a retention metric that will visibly get worse.

The diagnostic is the Asymmetry Ratio (Behavioral Gravity, frameworks/03, original framework, proposed here):

**Asymmetry Ratio** = *(steps to start) ÷ (steps to stop).* **I-03 obtains when the ratio ≈ 1.**

This is the hardest entry in Family I to evidence, and TAXONOMY.md open question 3 says so directly: "*Family I risks becoming a wish-list. Each entry needs a shipped, named product that actually does it, or it should be cut.*" This entry therefore leads with the evidence rather than the argument, and it states the limits of that evidence explicitly.

## PATTERN DNA

**Goal** acquire users who will not commit to a product they cannot leave, and retain them on the strength of the product rather than on the strength of the exit friction → **Trigger** the user's decision to leave, exactly as in F-09 and H-01. **Same trigger, opposite response.** That identity of trigger is what makes I-03 the clean counterfactual → **Interface** a cancel control in the same location as billing; one confirmation; an immediate effective cancellation; an email receipt with the effective date → **Bias** none is exploited. This is the row that distinguishes Family I from every other family in the taxonomy: **the Bias link is empty.** → **Emotion** relief, and, on reflection, trust → **Decision** the user leaves. Some of them come back → **Behavior** cancellation completes on the first attempt → **Reward** the user is not billed again, and they know it → **Business metric** this is the interesting one, and it is not "none." The metric is **willingness to sign up in the first place**, plus reactivation rate, plus the absence of chargebacks, complaints, app-store one-star reviews, and regulatory exposure. Family I is not charity (frameworks/02).

**The link that breaks on disclosure: none.** Full disclosure *strengthens* this pattern. "You can cancel in one click, any time, in the same place you signed up" is a sentence a company would print on its pricing page, and several do. That is the entire structural difference between an ethical pattern and a manipulative one, and I-03 is the cleanest demonstration of it in the taxonomy. Every other entry in

these nine has a link that snaps under narration. This one has a chain that gets stronger the more of it you say out loud.

### PSYCHOLOGICAL BASIS

The mechanism is **not** a bias, and this needs to be stated carefully, because a reader trained on the rest of this document will look for one.

1. **Trust as a precondition of commitment.** The claim is that a reversible commitment is easier to make than an irreversible one, and therefore that a visible, cheap exit *increases* entry. This is intuitive, it is the entire logic of a money-back guarantee, and it is the commercial case for I-03. **No empirical source for it was retrieved for this document.** CITATION NEEDED. It is stated here as a proposition, not as a finding, and a reader is entitled to be sceptical of it.
2. **The negative case is, by contrast, well-evidenced.** We know from Luguri & Strahilevitz (2021) that **obstruction works**: it was among the three strategies "particularly likely to manipulate consumers successfully" (Study 2, n = 3,777), and their broader finding was that under deceptive patterns "Decision architecture, not price, drove consumer purchasing decisions." This is important and uncomfortable for the argument this entry is making, so it is stated up front rather than buried: **removing exit friction will cost the company real, measurable, retained revenue.** Any claim that ethical design is simply free is contradicted by the best experimental evidence in the field.

That is why I-03 counts as a *repayment* of behavioral debt (frameworks/03) and not as a marketing exercise. The definition of repayment is that **the company voluntarily surrenders a metric it could have kept.** If it cost nothing, it would prove nothing.

1. **Sludge, inverted.** Thaler and Sunstein's "sludge" is friction that keeps people from what they *do* want. I-03 is the removal of sludge on a path where the company's interest lies in preserving it. Cited as a framework, not as an effect size.

**The honest summary of the psychological basis: the harm of the opposite pattern is empirically established, and the benefit of this one is not.** That asymmetry in the evidence base is real and is not concealed here.

### INTERFACE EXPRESSION

- A **"Cancel subscription"** control in the account or billing area, where a user looks for it, reachable in the same number of clicks as the subscribe control.
- **One** confirmation screen. It states what will happen, when, and what happens to the user's data. It does not argue.
- The cancellation takes effect on that click. The button tells the truth.
- An **email receipt** confirming cancellation and the final billing date. This is what makes the cancellation verifiable rather than merely asserted, and it is what makes the **H-01** Memory-axis cost (having to remember to check whether you actually cancelled) go to zero.
- **Same medium as signup.** Subscribed in the app, cancel in the app.

- Optionally: an offer, presented **once**, with "cancel anyway" visible on the same screen at equal weight (**I-02**). An offer is legitimate. A toll is not.
- Optionally: an exit survey, **after** the cancellation has completed, and optional.
- A stated data policy: what is kept, for how long, and how to get it out (**I-11**).

#### REAL-WORLD EXAMPLES

This is the section that decides whether the entry survives, so it is written to the evidence rather than to the thesis.

**37signals (Basecamp, HEY). The one named, documented, first-party commitment retrieved for this entry.**

37signals publishes a standing **Cancellation Policy** at 37signals.com/policies/cancellation. Retrieved , it states, in the company's own words, that "*we make it easy for you to cancel your account directly in all of our products*", and provides in-product, self-service cancellation instructions for each product (for HEY: the "Me" menu → "Account & Billing" → "Cancel your subscription"). The policy requires no phone call and no email exchange to cancel; conversely, and notably, it states that **an email or phone request to cancel is not itself a cancellation**, because the in-app mechanism is the mechanism. On data, it commits: "*We'll permanently delete the content in your account from our servers 30 days after cancellation, and from our backups within 60 days,*" giving a 30-day window in which the user may change their mind, and warning that content cannot be recovered once permanently deleted.

Three things make this a *usable* piece of evidence rather than a marketing claim, and each is independently checkable:

1. **It is a published, standing policy**, not an ad. It is a document the company can be held to.
2. **The policies are maintained in public version control** (the basecamp/policies repository on GitHub), so the *history* of the commitment is inspectable. A company that keeps its cancellation policy in a public git repository has made it expensive to quietly degrade.
3. **It states a genuinely unfavourable term honestly**: 37signals says it does **not** automatically pro-rate unused time, and invites the user to contact support. A pure-marketing exercise would have omitted that. Its presence is evidence that the document is a policy rather than a pitch.

**What this evidence does not establish, and the entry says so.** The policy documents *in-app, self-service, no-phone-call, no-questions cancellation*. It does **not** establish an exact step-for-step symmetry with signup, and this document has not counted the clicks on either path. **The claim made here is therefore the one the source supports: 37signals publicly commits to easy, self-service, in-product cancellation. The stronger claim, that the Asymmetry Ratio is exactly 1, is not asserted**, because it was not measured. Overstating it would be exactly the failure this project's Citation Law exists to prevent.

**Apple App Store subscriptions: a systemic instance.** Subscriptions purchased through the App Store are cancelled through the same account-level Settings surface on the same device, and this is a documented platform-wide mechanism rather than a per-developer choice. Its significance is structural: it demonstrates that **the exit path can be moved out of the merchant's control entirely**, which

removes the merchant's incentive to obstruct it, because the merchant no longer owns the screen. Where a platform intermediates cancellation, **H-01** becomes technically impossible for every app on it. This is worth naming as its own finding: **the most effective known remedy for a cancellation maze is not persuading merchants, it is taking the exit out of their hands.** CITATION NEEDED for a retrieved primary source describing Apple's subscription-cancellation flow; the mechanism is uncontroversially observable but no Apple support document was fetched for this entry.

**Amazon, by court order (2025).** The FTC's $2.5bn settlement with Amazon (25 September 2025) requires Amazon to present a **clear cancellation button** and to provide a **simple cancellation mechanism through the same medium the customer used to enrol.** This is I-03, imposed. It is included here as a boundary case with an important lesson: **an I-03 that is compelled is not a Family I pattern in the sense this taxonomy means.** Family I patterns are voluntary surrenders of a metric. A compelled surrender is a penalty. The distinction matters, because the taxonomy's normative argument depends on demonstrating that companies *can choose* this, and a court order demonstrates only that they can be made to.

**A candour note, and it is the most important sentence in this entry.** The task that produced this document asked for a named real product that actually does this, and required that a failure to find one be stated rather than papered over. One was found, and it is documented and first-party. But **one is a thin evidence base for a pattern that the taxonomy needs in order to make its normative argument work**, and the thinness is itself a finding: frictionless exit is *rare*, and it is rare for exactly the reason **H-01**'s Business Rationale section gives. The scarcity of examples is not a failure of this research. It is the result.

#### BUSINESS RATIONALE

The rational case, stated so that a growth executive would not laugh at it.

1. **The exit is priced into the entry.** A user deciding whether to start a subscription is implicitly pricing the difficulty of ending it. A visible, credible, cheap exit lowers the perceived cost of trying, and it is the same logic as a returns policy in retail, which is not run by retailers as an act of generosity. CITATION NEEDED for empirical support of this specific claim in software; it is a proposition, not a finding.
2. **Trapped revenue is the lowest-quality revenue on the P&L.** A user who failed to cancel is not engaged, will not expand, will not refer, will chargeback at a higher rate, will leave a one-star review, and will churn eventually anyway, with hostility. Retention obtained by obstruction shows up in exactly the same cell of the dashboard as retention obtained by value, and this is the core failure of standard instrumentation that Behavioral Debt (frameworks/03) is about.
3. **Honest instrumentation becomes possible.** A product with a frictionless exit gets a **clean churn signal**. Every cancellation is a real judgment about the product's value, uncontaminated by how many screens the user was willing to endure. That is a genuinely more useful number to run a company on, and it is unobtainable in a product with an **H-01**.
4. **The regulatory exposure goes to zero.** Post-Amazon, the expected cost of a cancellation maze includes a tail risk denominated in billions. I-03 is, among other things, the cheapest available insurance policy against **F-09**.

5. **It is a credible signal, precisely because it is costly.** Anyone can say their product is good. Only a company that believes it can afford to make leaving trivial. **The cost is the message**, which is the whole logic of signalling: a claim that costs nothing to make carries no information.

### BEHAVIORAL COST INDEX

Scored per frameworks/02. Reproduced from that framework's worked example.

| AXIS | SCORE | JUSTIFICATION |
|---|---|---|
| Time | 0 | Costs the user nothing beyond the action itself, which they intended to take. |
| Attention | 0 | Nothing to notice, nothing to guard against, no diversion to resist. |
| Memory | 0 | Nothing to retain. The email receipt means the user does not even have to remember to check whether it worked, which is precisely the H-01 Memory-3 cost, eliminated. |
| Trust | 0 | Spends none. **Builds it.** The BCI floor is 0 and cannot express a negative cost, which is a genuine limitation of the instrument and is noted here rather than hidden. |
| Money | 0 | None. |
| Privacy | 0 | None. |
| Emotion | 0 | None imposed. |

**Max = 0 → Ethical.**

Two observations about this row of zeroes, because a table of zeroes looks like a table that was not thought about.

**First, it is the *only* profile in the taxonomy that scores 0 on every axis**, and that is a substantive claim, not a shrug. Most ethical patterns still cost the user something: **B-01** Progressive Disclosure costs attention, **I-04** Usage Transparency costs a little emotion when the numbers are unflattering. I-03 costs the user nothing on any axis, because it is the removal of a cost rather than the imposition of a benefit.

**Second, the instrument's floor is its limitation.** The BCI measures the gap between cost paid and cost understood. It has no way to score a design that *returns* something to the user. A pattern that builds trust and a pattern that is merely trust-neutral both score 0 on the Trust axis, and they are not the same design. This is a real defect in the framework, it is inherited from the decision to score costs rather than net welfare, and frameworks/02 should record it in its Limitations section, where it currently does not appear.

### VALENCE ANALYSIS

**Ethical, and the argument has to be made rather than assumed, because "the good pattern is good" is not an argument.**

The valence rule in PROJECT.md §4 defines *ethical* as: the user's interest and the business interest align, and the user retains autonomy. Test both clauses.

**Autonomy: unambiguous.** The user can leave. That is the definition of autonomy in this context, and it is satisfied completely, not partially.

**Alignment: this is the clause that needs the work, and the honest answer is that alignment is *contingent*, not automatic.** The business interest served by I-03 is long-run: trust, willingness to sign up, clean churn data, no regulatory tail. The business interest served by **H-01** is immediate: this quarter's retained MRR. **These conflict, and the conflict is real.** Luguri & Strahilevitz establish that obstruction works. A company adopting I-03 is choosing a smaller certain number now over a larger uncertain number later, and it is entirely possible for a company to be *wrong* to do that, in the narrow financial sense, particularly a company with a short horizon, an exit to sell into, or a board that reads the retention chart monthly.

**So the honest statement of the valence is this.** I-03 is ethical *because* the user retains autonomy and pays nothing. It is *not* ethical because it happens to be profitable. The alignment claim is a hypothesis about long-run value, and frameworks/03 states the falsification condition explicitly: if products with high Opacity retain just as well over multi-year horizons, **the Behavioral Debt theory is wrong**, deceptive patterns are simply free, and the only argument left against them is a moral one. This document does not have the longitudinal data to settle that, and says so.

**Which leads to the sharpest thing this entry has to say.** Frictionless exit is the pattern on which the entire normative project of this taxonomy stands or falls. If it can be shown that companies which surrender exit friction are not punished for it, the whole Family I argument, that every harmful pattern's legitimate business goal has a non-harmful solution, becomes an empirical claim with support. If they *are* punished for it, then Family I is a moral appeal wearing a business case, and the document should say so plainly rather than pretend otherwise. **We do not currently know which.** That is the open question, and it is a more useful thing to end on than a reassurance.

#### ETHICAL ALTERNATIVE

Not applicable, and the template's insistence on this section is instructive here rather than awkward. **I-03 *is* the ethical alternative** to **F-09** and **H-01**, and Family I exists precisely so that the taxonomy's normative argument is made by *demonstration* rather than by condemnation (TAXONOMY.md, Family I preamble).

What can be said instead is where I-03 is **insufficient on its own**, because an exit that is easy but leaves the user's work behind is not really an exit:

- Pair with **I-11** **Portability by Design**: the user should leave with their data, in a usable format, without asking. 37signals' policy is partial here (it commits to *deletion* timelines, which is a different and lesser commitment than *export*).
- Pair with **I-08** **Reversible Commitment**: any commitment unwound on the terms it was made.
- Pair with **I-10** **Unused-Value Alert**: proactively telling a user they are paying for something they do not use. This is I-03's harder sibling, because it does not merely *permit* the exit, it *prompts* it. Very few companies do this, and it is the strongest possible signal a subscription business can send.

### REGULATORY STATUS

- **This pattern is not regulated, because it is what regulation is trying to produce.** It is the *remedy*, not the offence.
- **United States.** The FTC's settlement with Amazon (25 September 2025, $2.5bn under FTC Act §5(a) and ROSCA) imposes I-03 as injunctive relief: a clear cancel button, and cancellation in the same medium as enrolment.
- **United States, rulemaking.** The FTC's revised Negative Option Rule ("click-to-cancel"), which would have mandated I-03 for all online subscriptions, was **vacated in its entirety by the Eighth Circuit on 8 July 2025** on procedural grounds (no preliminary regulatory analysis, after an ALJ found the economic impact would exceed $100 million). The court did not hold the substance unlawful. The FTC has moved to revive it. **State automatic-renewal laws, which impose overlapping requirements, remain in force**, so a company shipping I-03 today is, in most US states, ahead of a requirement rather than beyond one.
- **European Union.** "Dark patterns" were codified into EU law in 2022 through the Digital Services Act, the Digital Markets Act, and the Data Act proposal. (Source: Gray et al. 2024.) The DSA's specific prohibition on deceptive interface design, article number and wording, was **not retrieved for this document**: CITATION NEEDED.
- **The strategic reading, for a company deciding today.** The legal floor is unstable and moving upward. The cost of building I-03 is one sprint. The cost of not having it, per Amazon, has a documented ceiling of $2.5 billion. A company that ships I-03 voluntarily converts a compliance liability into a marketing asset, which is the only circumstance in which "ethical design" and "commercially obvious" reliably coincide.

### RELATED PATTERNS

- **F-09** Roach Motel ATTR VERIFIED and **H-01** Cancellation Maze: the patterns this one negates. **The three entries are designed to be read together**; the Behavioral Gravity Asymmetry Ratio is the common instrument.
- **I-02** Symmetric Consent: the same equal-cost principle applied to consent rather than to exit. **G-01** is to **I-02** as **H-01** is to **I-03**, and the parallel is exact.
- **I-08** Reversible Commitment, **I-11** Portability by Design, **I-10** Unused-Value Alert: the patterns that complete the exit.
- **H-02** Exit Offer: legitimate at the boundary of I-03, and the boundary is precisely whether "cancel anyway" is visible on the same screen at equal weight.
- **H-04** Deletion Obstruction, **H-06** Data Hostage: the account-level harms that survive even a clean subscription cancellation, and which I-11 is needed to address.
- **D-12** Reputation Lock-in: the non-financial exit cost that I-03 does not touch at all.

### SOURCES

Verified project sources (see research/sources/):


- **Luguri, J., & Strahilevitz, L.J. (2021).** "Shining a Light on Dark Patterns." *Journal of Legal Analysis* 13(1), 43-109. DOI 10.1093/jla/laaa006. (Study 2 n = 3,777; **obstruction** among the three strategies

most likely to manipulate consumers successfully; "Decision architecture, not price, drove consumer purchasing decisions." Cited here **against** this entry's commercial case, as the best available evidence that removing exit friction has a real cost.)

- **Thaler, R.H., & Sunstein, C.R. (2021).** *Nudge: The Final Edition*. New Haven: Yale University Press. ("Sludge," the authors' own term for friction that keeps people from what they do want. Framework only; no effect size asserted.)
- **Gray, C.M., Santos, C.T., Bielova, N., & Mildner, T. (2024).** "An Ontology of Dark Patterns Knowledge." CHI '24. DOI 10.1145/3613904.3642436. (DSA / DMA / CPRA codification timeline.)
- **Mathur, A., et al. (2019).** "Dark Patterns at Scale." *PACM HCI* 3(CSCW), Art. 81. DOI 10.1145/3359183. ("Hard to Cancel," the sole type in their Obstruction category: the pattern I-03 negates.)

Sources:

- **37signals, Cancellation Policy.** https://37signals.com/policies/cancellation ("we make it easy for you to cancel your account directly in all of our products"; in-app self-service cancellation per product; an email or phone request is not itself a cancellation; "We'll permanently delete the content in your account from our servers 30 days after cancellation, and from our backups within 60 days"; no automatic pro-rating of unused time. **All quotations taken from the page.**)
- 37signals, Policies index. https://37signals.com/policies
- 37signals, basecamp/policies public repository. https://github.com/basecamp/policies (the policies are maintained in public version control, which makes the history of the commitment inspectable. This is the evidential point, not a claim about the policy's content.)
- 37signals / Basecamp help centre, billing and cancellation. https://help.37signals.com/billing/questions/316-information-about-adjusting-your-plan-or-canceling-your-account
- Katten, "FTC's Landmark $2.5 Billion Amazon Settlement Highlights Ongoing Focus on 'Dark Patterns'." https://quickreads.ext.katten.com/post/102l7e6/ftcs-landmark-2-5-billion-amazon-settlement-highlights-ongoing-focus-on-dark-p
- Goodwin, "FTC Settles with Amazon Over Enrollment and Cancellation Processes for Prime Memberships," September 2025. https://www.goodwinlaw.com/en/insights/blogs/2025/09/ftc-settles-with-amazon-over-enrollment-and-cancellation-processes-for-prime-memberships (the same-medium cancellation requirement and the clear-cancel-button requirement, which together constitute a court-ordered I-03).
- Cooley, "Click to Cancel Just Got Cancelled: Eighth Circuit Vacates Entirety of FTC's Negative Option Rule," 11 July 2025. https://www.cooley.com/news/insight/2025/2025-07-11-click-to-cancel-just-got-cancelled-eighth-circuit-vacates-entirety-of-ftcs-negative-option-rule
- WilmerHale, "Eighth Circuit Vacates the FTC's 'Click to Cancel' Rule, but Federal and State Regulators Likely to Remain Active," 1 August 2025. https://www.wilmerhale.com/en/insights/client-alerts/20250801-eighth-circuit-vacates-the-ftcs-click-to-cancel-rule-but-federal-and-state-regulators-likely-to-remain-active

**Outstanding verification tasks for this entry**, recorded rather than hidden:

1. No primary Apple support document describing App Store subscription cancellation was retrieved. CITATION NEEDED
2. The Asymmetry Ratio for 37signals (or any product) has **not been measured** by this research. The claim of exact step-symmetry is therefore **not made**.
3. No empirical source supports the proposition that a visible, cheap exit increases signups. CITATION NEEDED. It is the commercial case for the entire family and it is, at present, unevidenced.

APPARATUS

# Sources

A source appears here only if it was actually retrieved and read. Nothing is listed from memory. The "do not print" list at the foot exists because five confident, widely repeated claims turned out to be false.

**The single source of truth for citations in this document.**

A source appears here **only if it was actually retrieved and read.** Nothing is listed from memory. Where a detail could not be confirmed against the retrieved text, it is marked UNVERIFIED and must not be cited.

Fetch window: . Full retrieval notes, verbatim quotations, and per-source UNVERIFIED inventories live in research/sources/.

**Before citing anything from this file, read the "Do not print" list at the foot. It exists because five confident, widely repeated claims turned out to be false.**

---

### A. Deceptive design literature

Full notes: research/sources/01-dark-pattern-literature.md

**Brignull, H.** *Deceptive Design* (formerly Dark Patterns). Pattern library, 18 published types. Site began 2010 as darkpatterns.org; renamed on advice of the Tech Policy Design Lab. https://www.deceptive.design/types · https://www.deceptive.design/about-us

*Retrieved in full. The* **year 2010 is confirmed** *(corroborated independently by Gray 2018 and Gray 2024). The commonly cited day-level date of 28 July 2010 is* **UNVERIFIED***, it appears only in secondary sources. The original blog post was not retrieved.*

**Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L.** (2018). The Dark (Patterns) Side of UX Design. *CHI '18*, Paper 534. ACM. DOI: 10.1145/3173574.3174108

*Full PDF retrieved and read. Five strategies confirmed* **verbatim***: Nagging, Obstruction, Sneaking, Interface Interference, Forced Action. Corpus of* **118 artifacts***. Reproduces Brignull's 11 types as Table 1.*

**Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A.** (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *PACM HCI*, 3(CSCW), Article 81. DOI: 10.1145/3359183 · arXiv:1907.07032

*Abstract and project page retrieved. Confirmed:* **~11,000 sites, ~53,000 product pages, 1,818 instances, 15 types across 7 categories, 183 deceptive sites, 22 third parties.** *Per-type prevalence counts are* **UNVERIFIED** *(full body not read).*

**Luguri, J., & Strahilevitz, L. J.** (2021). Shining a Light on Dark Patterns. *Journal of Legal Analysis*, 13(1), 43–109. DOI: 10.1093/jla/laaa006

*Open-access full text retrieved.* **The empirical spine of this document.** *All figures quoted verbatim from the article body: Study 1* **n = 1,963***; Study 2* **n = 3,777***. Acceptance of a dubious service:* **11.3% control / 25.8% mild / 41.9% aggressive** *(the paper renders these as a 228% and 371% increase). Most-effective strategies: hidden information, trick questions, obstruction. "Must act now" urgency did* **not** *increase purchase. And the finding that carries the most weight for us: "Decision architecture, not price, drove consumer purchasing decisions." Per-strategy Study 2 percentages are* **UNVERIFIED***.*

**Gray, C. M., Santos, C. T., Bielova, N., & Mildner, T.** (2024). An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. *CHI '24*. DOI: 10.1145/3613904.3642436

**Cite the Crossref version of record, not the preprint.** *The Feb-2024 author PDF differs on both title and author order (it reads "…a Structure for Transdisciplinary Action," with Bielova before Santos). Same DOI. Verified against the Crossref registry directly. Confirmed: 10 taxonomies harmonised,* **64 types***, three levels. The 64 individual type names are* **UNVERIFIED** *(table not read).*

---

### B. Behavioral science

Full notes: research/sources/02-behavioral-science.md

**Fogg, B. J.** (2009). A Behavior Model for Persuasive Design. *Persuasive '09*, Article 40. ACM. DOI: 10.1145/1541948.1541999

*Full 7-page PDF retrieved and read.* **The paper contains no equation.** *"B = MAT" does not appear in it and must never be attributed to it. The model is stated in prose and two figures: behavior is a product of motivation, ability, and* **triggers***. Fogg renamed "trigger" to "prompt" in* **late 2017** *(confirmed verbatim at behaviormodel.org/prompts); the current form is B=MAP. The 2003 book is* **UNVERIFIED** *and is not the source of the model.*

**Ferster, C. B., & Skinner, B. F.** (1957). *Schedules of Reinforcement*. Appleton-Century-Crofts.

*Retrieved (Internet Archive scan).* **Cite both authors.** *"Skinner (1957)" is the most common citation error in persuasive-design writing. Book-level citation only: the interior was not read, so* **no response-rate or resistance-to-extinction number may be attached to it.** *Applying the operant literature to consumer apps is an* **analogy, not a finding***, label it as such.*

**Kahneman, D., & Tversky, A.** (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291.

*Scanned first page retrieved; journal header confirmed. Title is "decision* **under** *risk," not "decision making under risk," which several aggregators get wrong. Confirmed: the value function is "steeper for losses than for gains", the formal statement of loss aversion. The 1991 riskless-choice paper was* **not retrieved***; do not cite it.*

**Ruggeri, K., Alí, S., Berge, M. L., et al.** (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 4(6), 622–633. DOI: 10.1038/s41562-020-0886-x

*Retrieved.* **4,098 participants, 19 countries, 13 languages; replicated for 94% of items.** *Prospect theory is the sturdiest item in this bibliography. Cite it with confidence.*

**Nunes, J. C., & Drèze, X.** (2006). The Endowed Progress Effect: How Artificial Advancement Increases Effort. *Journal of Consumer Research*, 32(4), 504–512. DOI: 10.1086/500480

*Abstract retrieved verbatim. Confirmed: the* **8-step vs 10-step-with-2-granted** *manipulation, and that it raises both completion likelihood and speed. Mechanism is* **perceived progress, not sunk-cost avoidance***, routinely misreported.* **The car-wash "19% vs 34%" figures are UNVERIFIED and must never be printed.** *They do not appear in the abstract, which never mentions a car wash. Full text paywalled.*

**Ghibellini, R., & Meier, B.** (2025). Interruption, recall and resumption: a meta-analysis of the Zeigarnik and Ovsiankina effects. *Humanities and Social Sciences Communications*, 12, 962. DOI: 10.1057/s41599-025-05000-w

*Retrieved.* **The Zeigarnik memory effect does not replicate:** *pooled recall ratio* **0.99** *(k = 38). The authors conclude it "lacks universal validity."* **The Ovsiankina effect does:** *pooled task-resumption rate* **67.0%** *(k = 21). Progress bars and unfinished onboarding are* **resumption** *phenomena. Citing Zeigarnik for them is near-universal in UX writing and is wrong.*

**Thaler, R. H., & Sunstein, C. R.** (2021). *Nudge: The Final Edition*. Yale University Press. (Original edition 2008.)

*Publisher pages retrieved. "Choice architecture" is confirmed as* **their own coinage***. The Final Edition introduces* **"sludge"***, their term for friction that keeps people from what they do want, and the directly relevant concept for this document.* **No "nudge effect size" may be cited.** *The pooled effect is disputed under publication-bias correction. Cite the framework, or cite specific studies.*

**Cialdini, R. B.** (2021). *Influence, New and Expanded: The Psychology of Persuasion*. Harper Business.

*Publisher record retrieved.* **Seven principles, not six:** *Reciprocity, Scarcity, Authority, Consistency, Liking, Social Proof,* **Unity***. Confirmed on Cialdini's own site, which is internally inconsistent, still carrying a six-principle page.* **Replication is mixed and principle-dependent.** *Do not present these as uniformly settled science. First-edition year is* **UNVERIFIED** *(sources conflict, 1983 vs 1984).*

**Eyal, N., with Hoover, R.** (2014). *Hooked: How to Build Habit-Forming Products*. Portfolio/Penguin.

*Copyright page and table of contents retrieved. Four stages confirmed:* **Trigger, Action, Variable Reward, Investment.** *The byline credits Hoover.* **A practitioner framework with no primary empirical base of its own.** *Never cite it as though it were a tested theory.*

---

### C. Law and regulation

Full notes: research/sources/03-regulatory.md. All article numbers confirmed against primary text (EUR-Lex HTML, FTC filings, the Indian gazette), not secondary mirrors.

**FTC, Negative Option / "Click to Cancel."**

**THE RULE IS NOT IN FORCE.** *Vacated* **in its entirety** *by the Eighth Circuit in 2025, Custom Commc'ns, Inc. v. FTC, 142 F.4th 1060 (8th Cir. 2025), on procedural grounds (no preliminary regulatory analysis under FTC Act §22). The vacatur* **reinstated the 1973 Rule.** *The FTC restarted rulemaking with an ANPRM published* **13 March 2026** *(RIN 3084-AB54).* **No final rule exists.** *Any sentence beginning "US law now requires cancellation to be as easy as signup" is* **false.** *The case citation above was taken from footnote 40 of the FTC's own ANPRM.*

**EU Digital Services Act**, Reg. (EU) 2022/2065, **Article 25.**

*Full text retrieved from EUR-Lex.* **The phrase "dark patterns" appears in Recital 67, not in Article 25 itself.** *Art. 25(2)* **carves out** *anything already covered by the UCPD or GDPR, so most cookie-consent patterns are policed under* **GDPR, not the DSA. No confirmed Art. 25 fine exists as of July 2026.**

**GDPR**, Reg. (EU) 2016/679, **Arts. 4(11), 6(1)(a), 7**; Recitals 32, 42, 43.

**Art. 7(3) is the operative provision** *for exit design: "It shall be as easy to withdraw consent as to give it." This, not the vacated FTC rule, is the real legal basis for I-03 (Frictionless Exit).*

**EU Digital Markets Act**, Reg. (EU) 2022/1925, **Arts. 13(4), 13(6)** (anti-circumvention).

*No standalone dark-patterns article. Binds* **gatekeepers only.**

**EU Data Act**, Reg. (EU) 2023/2854, **Arts. 4(4), 6(2)(a)**; "dark patterns" named in Recital 38.

**California CPRA**, Cal. Civ. Code **§ 1798.140(l)** (definition), **§ 1798.140(h)**; **11 CCR § 7004(a)(2)** (symmetry in choice).

*Agreement obtained through dark patterns* **"does not constitute consent."** *CPPA Enforcement Advisory 2024-02:* **"Dark patterns are about effect, not intent."** *This is the Disclosure Test, encoded in statute.*

**India, CCPA Guidelines for Prevention and Regulation of Dark Patterns, 2023.** Notified **30 November 2023** under s.18, Consumer Protection Act 2019. **13 named patterns** (Annexure 1).

*All 13 confirmed against the gazette. Note* **#11 is "Trick Question,"** *not "Trick Wording" as several write-ups have it.* **TRAP:** *a widely circulated law-firm note renders the INR 20 lakh penalty as "~USD 24 million." It is approximately* **USD 24,000***, wrong by a factor of ~1,000. Do not copy it through.*

**OECD** (2022). *Dark commercial patterns.* OECD Digital Economy Papers **No. 336**, October 2022.

*Retrieved. Working definition quoted verbatim in the source file.*

**Enforcement actions with confirmed figures**

**FTC v. Epic Games (Fortnite):** *the widely cited "$520M dark-patterns fine" is* **two separate settlements***,* **$275M** *COPPA civil penalty* **plus $245M** *in refunds for deceptive patterns and billing.* **Only the $245M is the deceptive-patterns component.** *Citing $520M overstates it by more than double.* **FTC v. Amazon (Prime cancellation), 25 Sep 2025: $1bn penalty + $1.5bn redress = $2.5bn.** *FTC Act + ROSCA.* **UNVERIFIED, do not cite:** *CNIL Google/Facebook name-to-amount mapping (CNIL has depublished the release and anonymised the register; only secondary reporting remains). Irish DPC/Meta and Amazon Luxembourg figures were* **not retrieved** *and* **no number is given for them at all.**

---

**Do not print**

Five claims that are confident, widely repeated, and wrong. Each was caught by retrieval. Each would have survived an unaided read-through of this document.

| CLAIM | REALITY |
| --- | --- |
| "Fogg's equation, B = MAT" | **No equation exists in Fogg (2009).** Community retrofit. Also, "trigger" became "prompt" in late 2017. |
| The Zeigarnik effect (interrupted tasks better remembered) | **Does not replicate.** Pooled recall ratio 0.99. Use **Ovsiankina** (resumption, 67%). |
| Nunes & Drèze, "19% vs 34%" car wash | **Not confirmed against the primary text**, which never mentions a car wash. |
| "US law requires easy cancellation" (FTC click-to-cancel) | **Vacated by the 8th Circuit in 2025.** No final rule exists. Cite **GDPR Art. 7(3)** instead. |
| Epic's "$520M dark-patterns fine" | **$245M** is the deceptive-patterns half. The other $275M is COPPA. |

Plus: never cite a **nudge effect size**; never write **"Skinner (1957)"** without Ferster; never cite the **Kahneman & Tversky 1991** paper (not retrieved); never cite **Cialdini** as settled science without noting that replication is principle-dependent.

JOIN THE RESEARCH

# Reach out, push back, help build this

This book makes one promise: it can be checked. If a citation does not hold or a claim reaches past its evidence, that is the most useful thing you can send. Refutation is welcome and it is credited. And if you build consumer software and want to help run the audit, get in touch. This is meant to be built in the open.

**Get new chapters and the audit results by email.** The PDF stays free either way. No spam, unsubscribe in one click.

Subscribe for new chapters at amoghtiwari.com


**X / Twitter**
@itsamoghtiwari

**GitHub**
amoghtiwari34

**LinkedIn**
in/itsamoghtiwari

**Email**
amogh@beomelo.com


If reading the whole thing was worth your time, you can buy me a coffee. Entirely optional, and it keeps the research going.