The Three-Layer Trick

Or: Why You Should Never Show Anyone Their Score

Here's a fun thing about humans: the second you put a number in front of one and tell them "this number is you," they will spend the rest of their life trying to make that number bigger. They will lie. They will cheat. They will optimize their entire existence around the digits, even when the digits stop measuring anything real. Especially when the digits stop measuring anything real.

We have decades of evidence on this. Credit scores were supposed to measure creditworthiness. Now they measure how good you are at gaming credit scores. Klout was supposed to measure social influence. It became a parody so absolute that the company quietly dissolved itself in shame. Steps trackers were supposed to measure activity. Now there are people who shake their phone in their hand for an hour to hit 10,000. China built a literal social credit system and accidentally invented a new flavor of dystopia. Every productivity app I've ever installed turned me into someone who optimizes for the dashboard instead of the work.

This is Goodhart's Law, dressed up in a thousand outfits, all of them unflattering: when a measure becomes a target, it stops being a good measure.

So here's the question every reputation system has to answer, and almost none of them do: **how do you reward good behavior without telling people what the rewarded behavior is?**

Because the second you tell them, they're not doing the behavior anymore. They're doing the *measurement* of the behavior. And those are different things, even when they look identical.

The standard answer, and why it's stupid

The usual response to Goodhart is "hide the metric." Make it proprietary. Make it complicated. Don't show users their score, just show them outcomes.

This is fine until someone reverse-engineers it, and someone always reverse-engineers it. There are entire industries built around reverse-engineering the proprietary algorithms of search engines, recommendation systems, dating apps, and fraud-detection software. If your defense is obscurity, you have no defense. You have a slightly delayed surrender.

The deeper problem with hiding the metric is that you're still treating it like there's *one* metric. One number, hidden. One target, eventually reverse-engineered. One Goodhart waiting to happen, just postponed.

What if there were two?

What if the metric you showed users was deliberately, structurally, on purpose **not the same** as the metric the system actually rewarded?

That's the trick. That's the whole trick. Everything else is engineering.

Three layers

The Extropy Engine has three layers, and they have to stay separate or the whole thing collapses into another credit score.

The user-facing layer is what people see. Discounts. Money saved. A character sheet they curate themselves. Streaks if they want streaks. Levels if they want levels. Achievements if they want achievements. Whatever pulls people in. This layer can be as gamified as a Saturday morning cartoon. Doesn't matter. It isn't where the real scoring happens.

The merchant-facing layer is what businesses see. A free point-of-sale, a customer pipeline, operational data better than the KPIs they were paying consultants to interpret. Merchants don't have to believe in entropy reduction or thermodynamic ethics or any of this stuff. They just have to notice that customers are walking in asking if they're on the network, and that the POS is free, and that the merchant fees are 1.8% instead of 2.8%. They opt in for the same boring reasons businesses opt in to anything: it makes them money.

The engine is what nobody sees. The math. The validators. The decentralized ledger. The actual scoring function that produces actual EP that produces actual savings. Six tokens, one formula, one invariant: same entropy reduction equals same XP, regardless of who's reporting it. No reputation laundering, no Goodhart, no engagement farming.

The genius (and yes, I'm allowed to call my own work genius, this is my book) is that **Layer 1 and Layer 3 are deliberately not the same metric**. The user can grind streaks all day. Streaks pay out in dopamine. They don't pay out in EP. EP comes from the entropy-reduction signature underneath, which the user can't see, can't directly target, and even if they figured out the formula, validators would catch them faking the inputs.

It's not security through obscurity. It's security through **intentional metric divergence**. The thing you're rewarded for and the thing you're shown are correlated by design but they are not the same thing. You can't farm the real metric by farming the visible one. You can only farm the real metric by actually doing the thing.

Which, conveniently, is the entire point.

The character sheet thing

I keep using the Dungeons and Dragons metaphor because it actually works.

Your character sheet in D&D is yours. You filled it out. You named the character. You picked the class. You decide which parts to share with the table. You can keep a secret backstory the DM doesn't know about. You can flavor your stats however you want.

But you can't fudge the numbers. Not really. Not without everyone at the table noticing. The dice are the dice. The rules are the rules. If you say your wizard has 47 strength, the DM is going to ask some questions, and your wizard is going to fail a check, and the table is going to politely correct you.

Your sheet is yours. Reality is reality. Both things at once.

That's the system we're building. You hold the keys to your own record. You self-curate. You decide what's visible to whom. The state can't write to your sheet. Your employer can't write to your sheet. A foreign government can't downgrade you for posting the wrong meme. Only validated reality can write to your sheet, and only you can choose which parts of it to reveal.

You hold the pen and the eraser. Reality holds the dice.

That's the elevator pitch. That's the difference between this and every dystopian scoring system you've ever read about. In a social credit system, the state holds the pen. In a credit score, the bureaus hold the pen. In a Klout score, a startup with no idea what it's doing holds the pen. In ours, you do. The system can only witness. It cannot author.

What the merchant gets, in case you're wondering

Because honestly, this whole thing falls apart if merchants don't bite. The math could be perfect, the user experience could be sublime, and if no business signs up to honor the discounts, you have a beautiful loyalty program with nowhere to redeem.

So merchants need a real reason. Three of them, in order of how much they care:

1. **Customers are asking.** Once enough users are on the network, walking into a shop and going "are you on Extropy?" becomes a thing. Merchants will get tired of saying "no" and losing the sale. This is the same dynamic that got every coffee shop on every payment app eventually, except instead of just a payment, it's a customer pipeline.
2. **The POS is free.** Or close to it. Merchants are currently paying $80 a month to Square or Toast for software that hasn't been meaningfully updated in five years. We give them a better one for nothing and capture the merchant-services fee they were already paying anyway. Net: they save money. Net for us: we capture revenue. Net for the user: they save money. Everybody wins, except the incumbents, which is fine.
3. **The data is actually useful.** Most merchant analytics are vanity metrics. Daily revenue, top sellers, busy hours. Yawn. The Extropy ledger gives them entropy-reduction patterns across their customer base, which translates to: which products actually drive habitual loyalty, which products are bought once and never again, which time-of-day patterns predict customer retention. Things that show up nowhere on a Square dashboard.

By the time a merchant figures out what they're looking at, they're three years deep into the network and have nine of their suppliers also on the network and they're registered as a DFAO node and they're an infrastructure participant in something they originally signed up for to get a slightly cheaper credit card processor.

That's not a bug. That's the design.

The two-register thing

A note on language, because this matters.

If you've made it this far in the book, you can tell I curse. A lot. I'm fine with this. The sentiment behind "unfucking the world" is load-bearing for me. It's the title. It's the project. It's the thing.

But here's the part where I have to be honest with myself: the moment Extropy Engine has a customer-facing app on the App Store, and that app says "unfuck your finances," half the addressable market noped right out. Older small business owners. Suburban moms shopping at the local market. Church groups. School PTAs. The exact people whose participation makes the network actually work, because they're the ones with consistent habits and durable patterns.

So here's the deal I'm making with myself, and I'm telling you about it so you can hold me to it: **two registers, one project.**

Internally, in the repo, in the manifesto, in this book, in conversations with collaborators who get it, the language stays sharp. Unfuck the world. Unfuck the economy. Unfuck the loyalty program. That energy is what got me here and I'm not throwing it away to please a venture capitalist who needs everything to sound like McKinsey.

Externally, in the app, on the merchant page, in the consumer marketing, the language softens. "Rewire the everyday economy." "Fix the system from underneath." "Reward what actually matters." Same substance. Different packaging. The PTA mom doesn't need to hear me curse to want a 12% discount on groceries.

This is not selling out. This is recognizing that audience targeting is real and that the goal is to actually reach people, not to perform authenticity to an audience of zero.

If anyone reading this in five years catches me mixing the registers, please email me and yell at me. I deserve it.

The privacy tax problem

One last thing, because I want to flag it for anyone helping build this.

The whole "voluntary self-reporting" model has a failure mode, and it's the same failure mode every loyalty program has eventually fallen into: rewarding disclosure volume creates pressure to over-share, which means privacy-conscious users are effectively paying a privacy tax to participate, which means the system slowly selects for the people willing to surveil themselves the most, which means it becomes a social credit system through emergent behavior even though nobody designed it that way.

I don't want that. So the math has to be structured so it doesn't happen.

The fix is in the formula. R is per-domain rarity, a property of the action class, not the disclosure. ΔS is the entropy delta of a specific loop, not a function of how much else you've shared. F is frequency-of-decay, which actually penalizes over-reporting common actions because they decay fast. The math, by structure, does not reward you more for telling it more. It rewards you for *doing* more. Honestly.

That distinction has to be defended forever. The day someone proposes a feature that changes EP based on disclosure volume rather than action volume is the day this turns into Klout with a thermodynamics aesthetic, and we have to say no.

I'm putting this in writing in a published book so future contributors can quote it back at me if I forget.

The point

Most reward systems are designed by people who think the goal is to maximize engagement. They are wrong. The goal is to maximize the durable, honest behavior that the engagement is supposedly measuring. The engagement is the proxy. The proxy is not the thing. If you optimize for the proxy, you destroy the thing.

The Extropy Engine optimizes for the thing.

The user sees savings. The merchant sees customers. The engine sees entropy. Three layers, three views, one underlying coordination function that everyone benefits from without anyone having to understand it.

That's the trick.

You hold the pen. You hold the eraser. Reality holds the dice. Everybody saves money. The world unfucks itself, slowly, one validated loop at a time.

I'll allow myself one curse word in the public version. Just that one.