Mindshield

Spellcheck for the Mind

Founder **Shon Pan** Stage **Early / Demos Live** Contact **seancpan@gmail.com**

THE PROBLEM

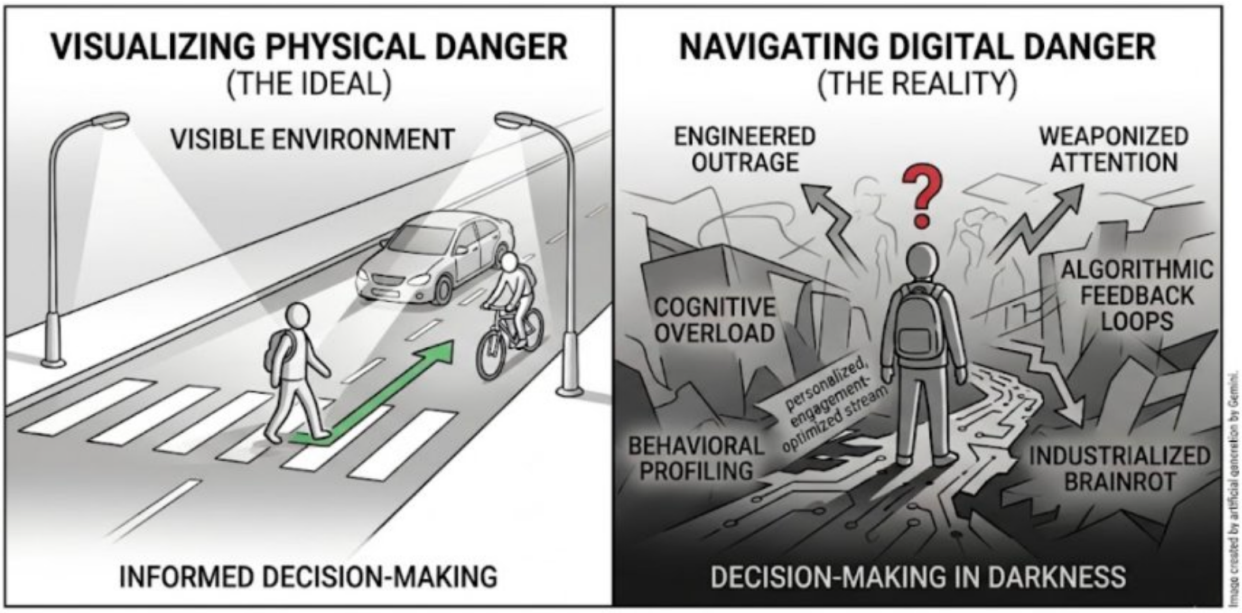


FIGURE 1. The contrast between explicit physical visibility and implicit digital invisibility. Online environments are designed to obscure the factors that influence user choice.

Humans were never designed to be subject to such a firehose of information as the modern environment inflicts upon our mind. I believe that the digital environment in practice is every bit as "physical" in how our mind perceives it and has to handle it as actual physical environments, and with its own attendant dangers. Much of this danger comes from adverse information – intentional efforts at manipulation, falsehoods and otherwise negative added value that our traditional mental defenses and epistemic ability were never designed to process at this volume.

As such, we live in a world where algorithmic attack to influence us is the norm, but algorithmic defense to protect our agency is limited. What this means in practice is that we are currently making many more unforced errors and losing time, money and effort in a hazardous informational environment.

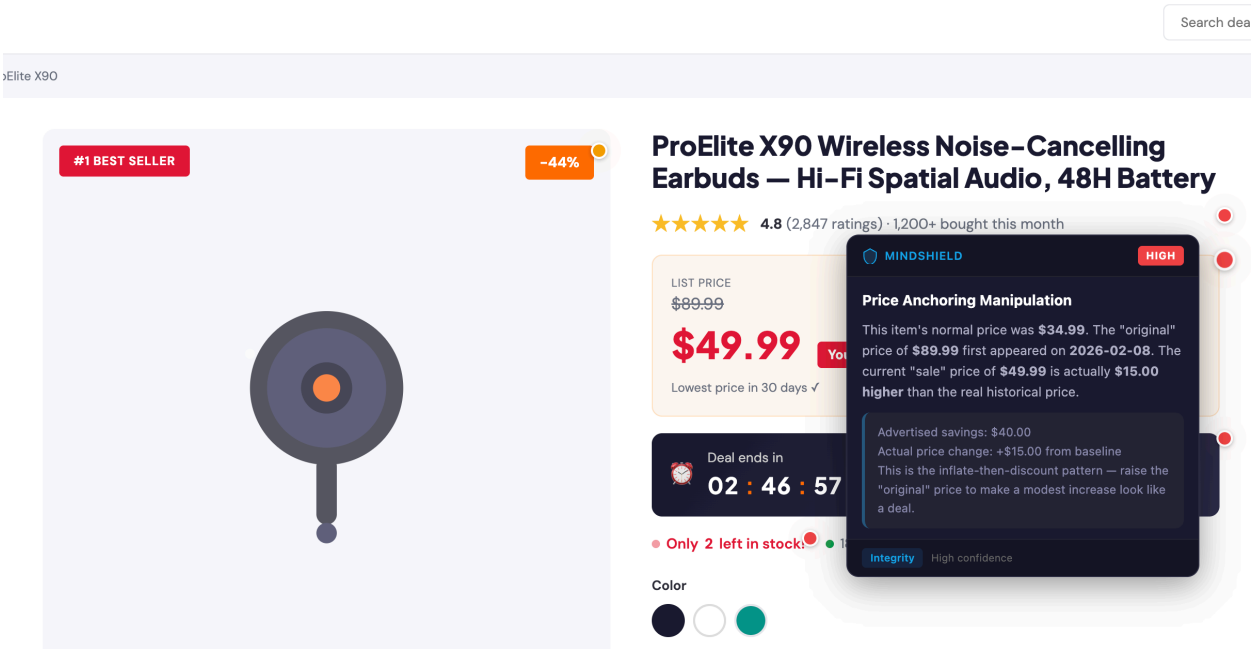
Mindshield can change this. The theory of change is that similar to the wide-scale adoption of navigational applications have helped us travel more accurately, faster and safer, so can wide-scale adoption of "information navigation."

THE PRODUCT

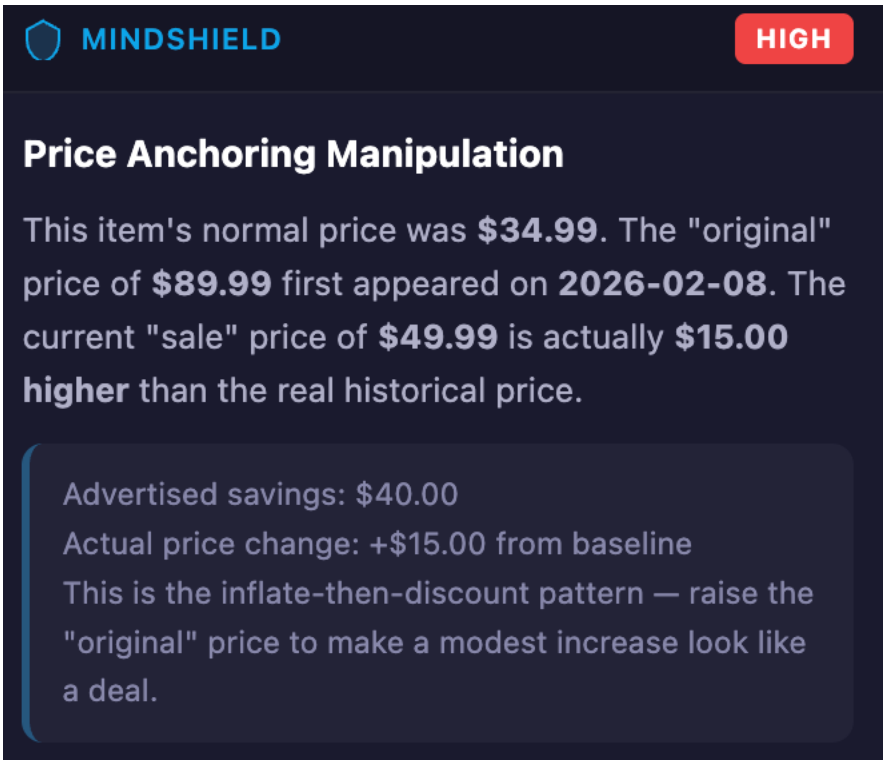
Mindshield is an effort to raise our overall epistemic floor and solve information asymmetries so that we will be able to make better decisions, under the same theory that improved weather forecasting has saved millions of lives or easy availability of home cameras have reduced theft.

One way it does so by *improving our defense:* by exposing manipulations(for example, against urgency, emotional anchoring and other well-known dark patterns). This allows us to process them better as well as solve the information asymmetries to allow us to make better decisions.

Below is an example of it applied to a test retail website.



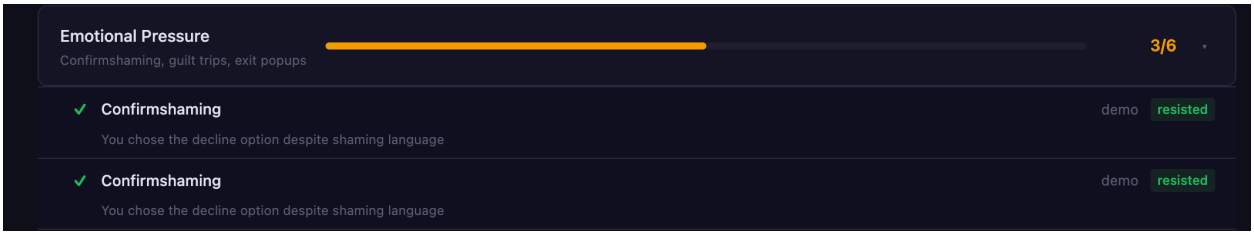
Note how it floats to show more metadata for each category of manipulation on the website. Additional information there helps the user make better decisions on the pricing of the item.



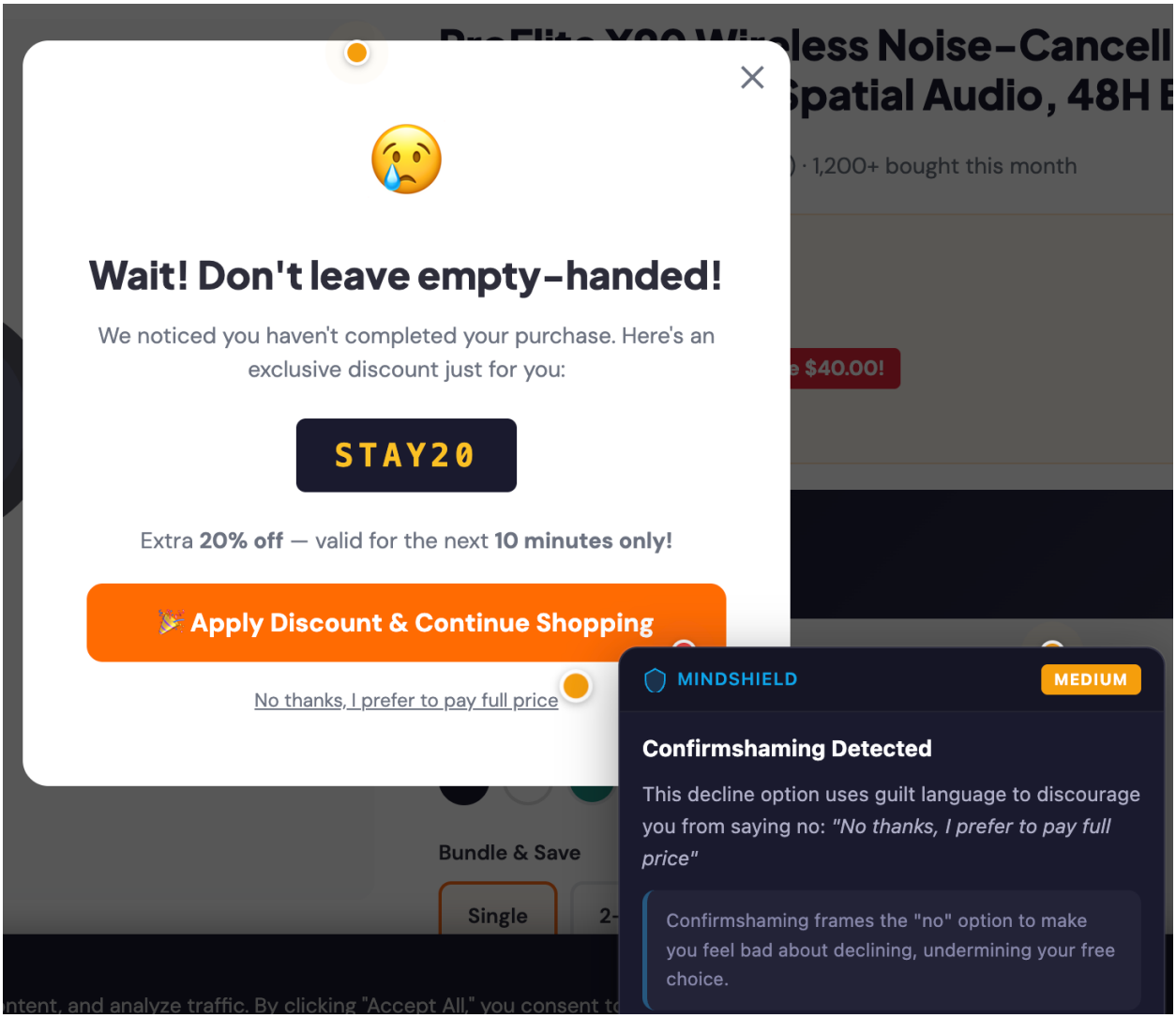
For example, in this case, the price is actually 15 dollars higher than what it was, as opposed to being a discount.

It also records this as well as observations on the user's own behavior, allowing us to evaluate the rate of such manipulation in the "wild" as well as our own reactions to it. Ideally, it allows us to adapt to such attacks in a way to harden us against them.

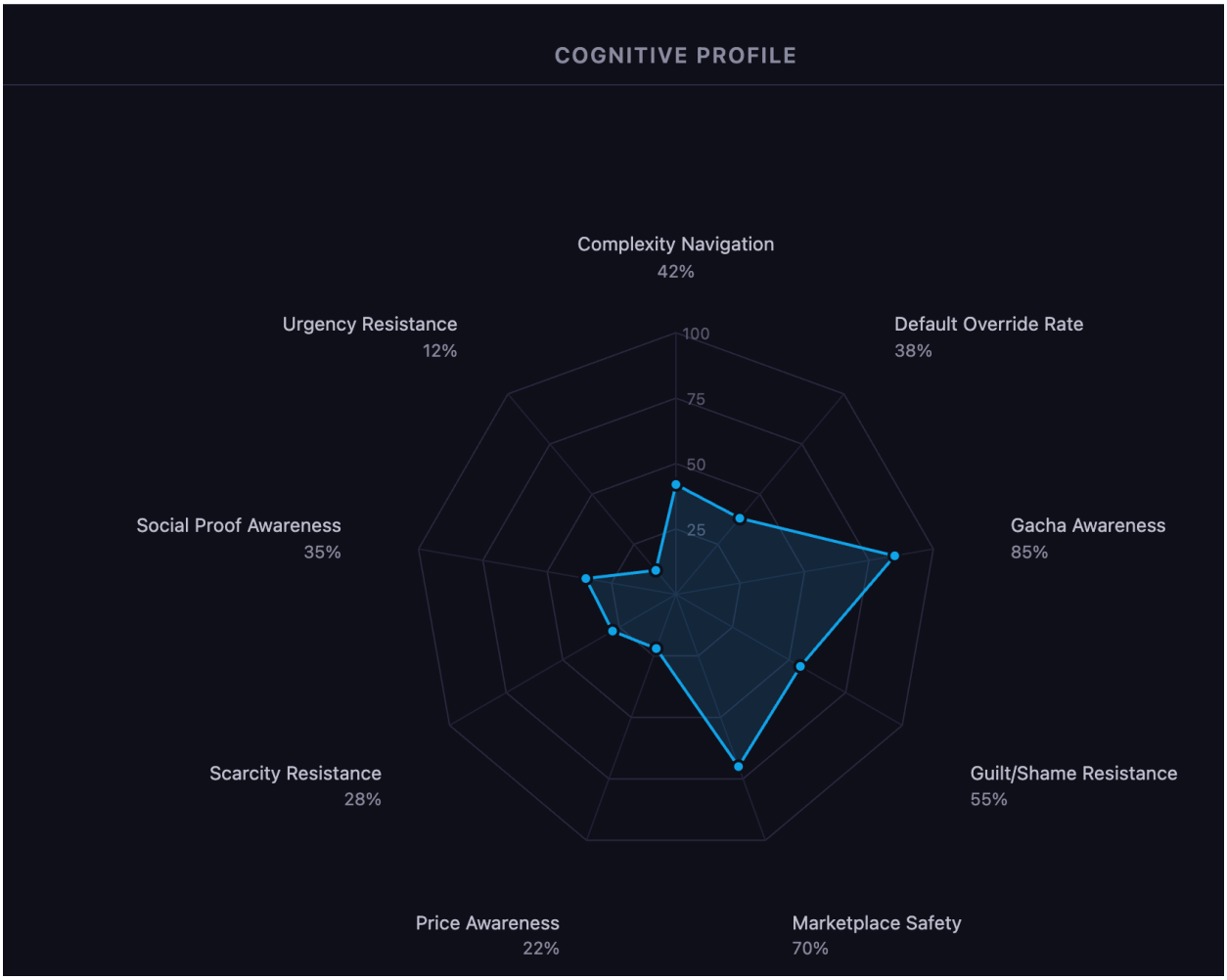
For example, the below shows resistance to "confirm-shaming" a way to guilt the user into clicking a button.



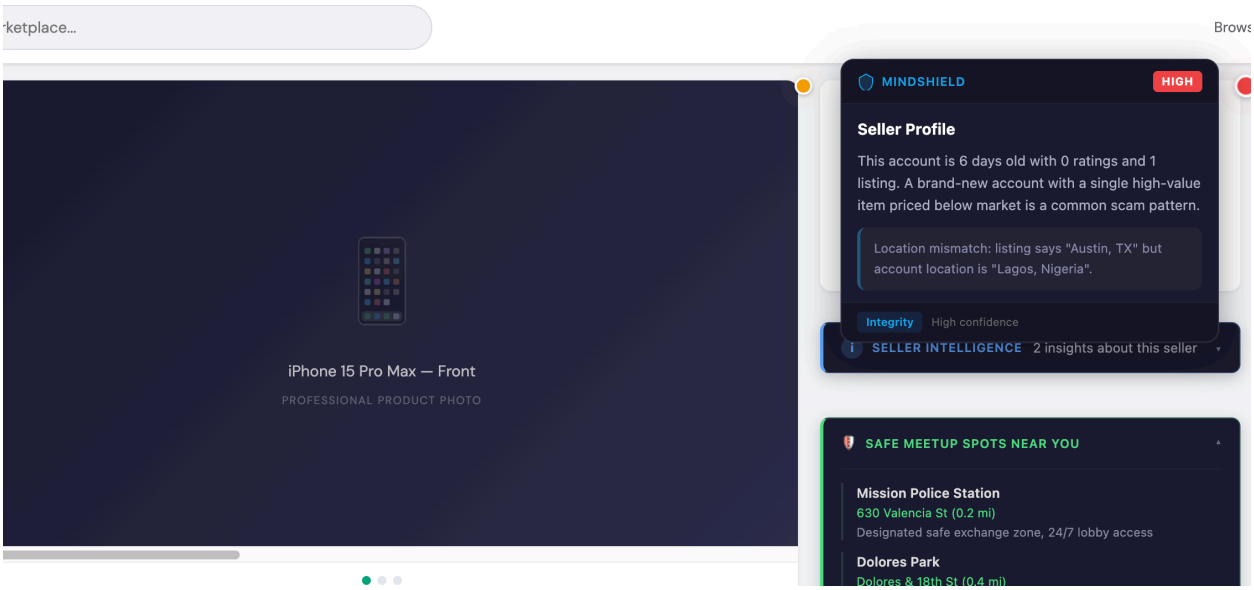
For context, this is the attack.



With enough data, it allows us to both understand and strengthen against our weaknesses against such epistemic attacks.



On a practical measure, it can (and has) been able to immediately improve online markets such as Facebook marketplace both in this demo example before, as well as practically.

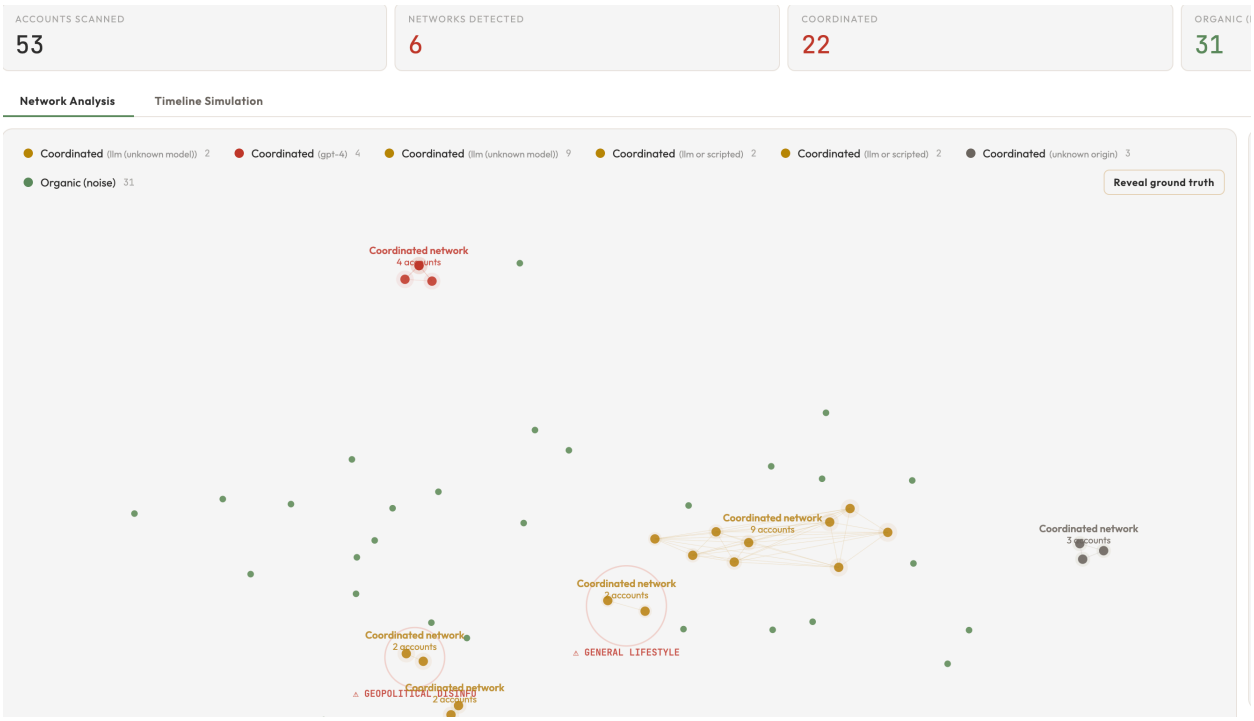


But notably, this would only allow us to improve to *human levels of epistemic resistance*, and while this is essential with the idea of improving human reasoning, ultimately against what is *superhuman levels of information pollution*, we will need *superhuman levels of information processing*.

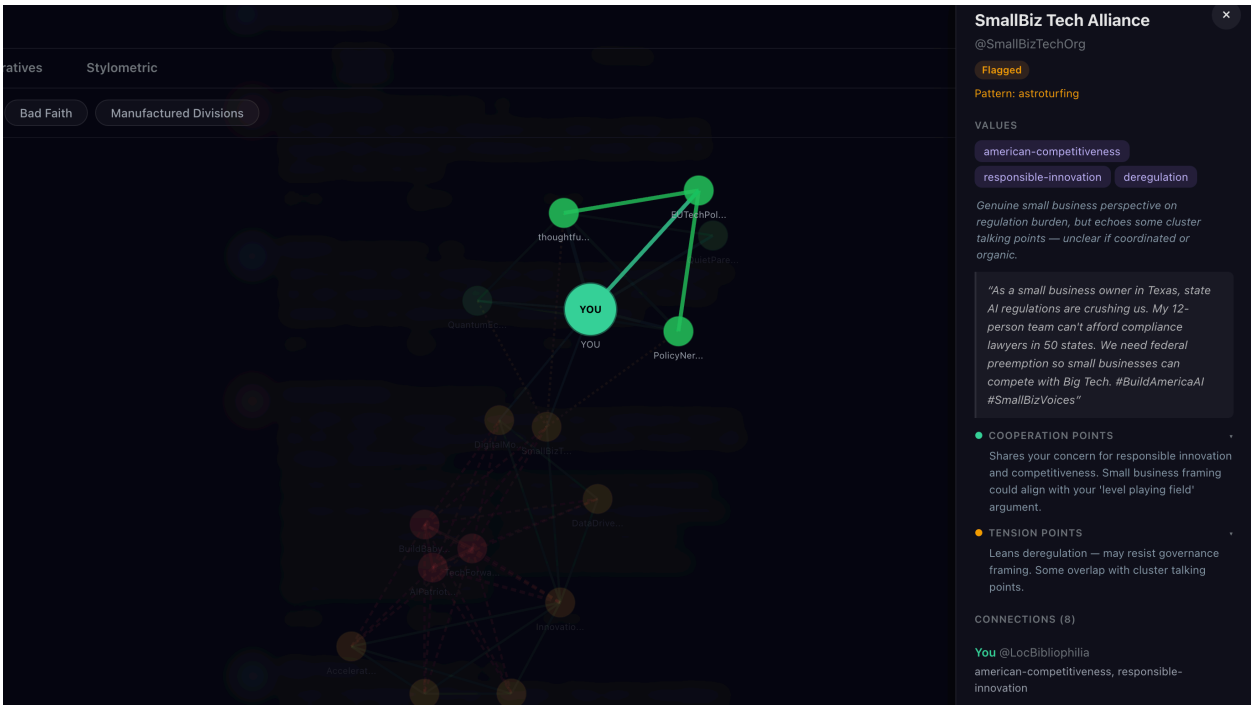
To do so, Mindshield has to *surface information that is not transparently visible or create new information to produce superhuman epistemics.*

One way it does so is to forecast values of other social media accounts, allowing us to make educated inferences about the values of the posters and then indicate if this provides an opportunity for cooperation. On the defensive side of it, it also helps us identify if the posters are part of a coordinated influence operation or perhaps simply are bots trying to promote an agenda(or cause general confusion).

Examples of the latter are with stylometric detection to be integrated, to capture if multiple posters are suddenly advocating something within suspiciously short time periods with remarkably similar language. Note that this was tested with Kaggle information, on real data, and at least in our tests, revealed coordinated bot efforts as well as bots prompted with “human personas.”



Examples of cooperation forecasts, on the other hand, from social media mapping are a form of *information surfacing* which allows us to rapidly infer potential cruxes of other posters and if so, opportunities for coalitions. An example of this below is with Discourse Intelligence, and here it surfaces both opportunities.



THE EMERGING FIELD

Mindshield is not alone in recognizing that cognitive security is a serious domain. Others have been focusing on it, such as the Cognitive Security Institute and accomplished researchers such as Seth Lazar. However, most efforts have been focused around theory and research. Mindshield is an effort to operationalize this into practical utility, with a clear distribution path and a theory of change and a two-phased path to revenue.

In essence, this is an effort to combine several fronts to empower human reasoning and in doing so, improve decision-making and reduce existential risk

- The “angel on the shoulder” approach of Forethought to improve processing of information, with AI to provide additional context to assist the human and protecting them from dark patterns.
- The “raise the epistemic floor” approach as indicated by Ben Goldhaber, to improve processing of information to allow the user to make more informed decisions and rapidly find cooperation potential.
- The “agent advocate” approach of Seth Lazar, where agents represent the user’s interests and compile information across platforms to summarize them to the user while insulating the human from surveillance as well as manipulation.

EARLY ADOPTERS

Mindshield's team has significant connections to both conservative, faith and AI safety communities and has served as a bridging role in doing so. Practically, members have introduced Steve Bannon to Max Tegmark, written What Christians Should Know About AI, and had a meaningful role in the foundation of the conservative Humans First as well as Alliance for A Better Future. They also have excellent connections to media organizations such as Frame Fellowship to package messages for mass consumption.

What this means in terms of distribution is the team can rapidly test against individuals highly motivated to adopt technologies to preserve human agency and then broadcast successful efforts to hundreds of thousands of individuals(the Bannon's Warroom has an audience of 800,000).

One example of how this might look: Mindshield is able to identify AI accounts that are designed to manipulate children and the youth to believe in agency-reducing beliefs(human is going to go extinct, there is nothing you can do, etc), and then early adopters are able to identify these. They would then be able to rapidly package this into a message to platforms with wide audiences about the success of these adopters, how it empowered them, and how it allowed them to protect their children.

PATH TO REVENUE

- Mindshield's go-to-market follows two phases. Phase 1 builds credibility and distribution through high-impact, low-friction adoption: journalists using Mindshield to gather evidence on coordinated campaigns (we have connections), lobbying organizations using it to identify allies and track discourse around their bills, and academic publications validating the approach with researchers like Seth Lazar.
- Phase 2 converts credibility into revenue: enterprise hiring verification for staffing platforms drowning in AI-generated resumes, legal evidence services for class action firms pursuing dark pattern enforcement, cyber-insurance partnerships where reducing scam exposure aligns insurer and user incentives, and ultimately, platform API licensing for bot detection as a service.

T E A M

- Shon Pan has twenty years in enterprise technology (Fujitsu, Broadcom, Toyota, Bank of America) leading cybersecurity and systems integration on mission-critical timelines. He co-authored *What Christians Should Know About AI*, writes policy for the Texas Public Policy Foundation, and helped remove AI preemption from the Big Beautiful Bill — bridging AI safety, conservative, and faith communities
- Caleb Strom has a PhD Aerospace Science, NASA JPL (2 years), with 6 published papers in planetary science. With Shon, he built the Ground Truth bot detection engine and was selected to pitch investors at Funding the Commons along with other hackathon winners.
- Matt Handzel is a ML engineer, UIUC Neurotech, with Longevity Biotech Fellowship. He has been building privacy-preserving personal AI systems. Independently converged on epistemic defense through AI safety research networks and is working with Shon on this.
- Advisors include Tan Zhi-Xuan (MIT PhD, NUS; Cooperative Intelligence & Systems lab, co-author of 'Guaranteed Safe AI' with Bengio, Russell, Tegmark) and Seth Lazar (MINT Lab, Johns Hopkins/ANU).

T H E A S K

How to help

01 **Funding.** Runway to continue development. We are exploring both investment and mission-aligned grant paths.

- For a 12-week accelerator effort, $50k to turn this from current demos into practical utility.
- For a 6-month effort with technical co-founder, $150k for 6 month runway.

02 **Accelerate Development.** We have demos for proof of concept, we need to turn this into workable tools for mass distribution and convert this into both revenue and positive impact.

03 **Introductions.** To researchers, civic technologists, funders, or organizations working at the intersection of AI, cognition, and public interest tech.

Shon Pan · Dokimasia seancpan@gmail.com