Surrogation

Table of Contents

Surrogation is the substitution of a representation—be it a metonym, symbol, proxy, metric, or signal—for some represented whole.

1. Introduction

Goodhart's Law, in the (re)phrasing of anthropologist Marilyn Strathern: "When a measure becomes a target, it ceases to be a good measure." Since we expect humans to behave roughly rationally toward self-interest, and as the incentive structure of an activity *determines* self-interest, the surrogate measure distorts behavior in the direction of optimizing for the *surrogate*, at cost to the surrogated which is proportional to the divergence *between* surrogate and surrogated.

We will call this general mechanism—where a *representation* of a holistic target generates its own gravitational field, and in some meaningful way

replaces that original whole—"surrogation." Choi, Hecht, and Tayler in the early 2010s published papers in management accounting theory (2011, 2012) proposing that managers exhibited a pattern of losing sight of their original strategic target in favor of an instituted proxy measure set up to represent it. The authors are potentially the first to use the term "surrogation" in this context, and they include in their definition the psychological "amnesia" of managers—but it is well-chosen as an umbrella handle for a broader phenomenon comprising multiple theoretic carvings. In management studies, that of Choi, Hecht, and Taylor. In the social sciences, that of Goodhart as well Law and that of Donald Campbell (i.e. "Campbell's Law") stating that "[t]he more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (In the context of policing, specifically, Campbell has accused the Nixon administration's crackdown on crime as having "as its main effect the corruption of crime-rate indicators, achieved through underrecording and downgrading the crimes to less serious offenses.") In philosophy, we see C. Thi Nguyen's theory of value capture as outlined in 2020's *Games and the Art of Agency* (and presented in the larger context of *gamification*, a process similar to surrogation). Nguyen defines *value capture* as the substitution of a simplified metric, or indicator, for a richer holistic value, distinguishing it from Goodhart's Law in that the substitution is *internalized* by the agents situated within the surrogate incentive structure. There is not just a change in the agents' behavior but in their actually held values.

The surrogate by definition is chosen because it is somehow easier or more tractable than the surrogated "real" or "original" destination. That previous destination may have beeen hidden from sight; it may have been too difficult to compare or rank among instances; it may simply have beeen costlier to track in time or money. A proxy is a surrogate; so is a signal, a metric, a marker, and a representation.

Surrogation is linked in meaningful ways to the concepts of "economic thinking" and "commodification," where the formalization, compression, or technical specification of a vague or humanistic value is "lossy," i.e. cannot meaningfully capture the whole. You can see its economically reductive—its "classical"—form in an especially weak passage of Brian Eno's 2015 lecture, "An Ecology of Culture" (emphasis mine):

> I started wondering about the genesis of that term ["the music industry"]. I can sort of understand why it's used because, you know, people who work in the creative arts are always desperate to try to get a little bit of money from government and apparently the way ofconvincing them that they should give you some money is to tell them that you're an industry. If you're an industry that means you're part of the economic framework and *that everything you do can ultimately be expressed as a single number*. Like your contribution to GNP or the number of jobs that you provide or things like that, the number of Number Ones you've had.

But surrogation is not a problem or phenomenon unique to metrics. Setting reality to words is always the first surrogation: we reify our concepts, confuse them with nature. Were we to avoid quantitative analysis, we would not avoid surrogation. Language, like our internal drives and desires, is always vague—our goals are always underspecified, and our words are always unstable and underdefined. One way to put *surrogation* in simple English, losing the unnecessary if common statistical bent, is to say that some simplified marker of *appearance*, eligible for its co-occurence with an intractable or invisible *reality*, becomes in some way a substitute *for* that reality.[1] In the language of signaling theory, the sign is reified in place of the hidden, signaled quality. Markers in fashion—an everday, material instantiation of signaling theory—are famously contextual. The same piece of clothing can signify very different aspects of its wearer depending on the larger inferred complex of intentionality, knowingness, and providence in which its display (is inferred to) originate. A “designer” brand like Lacoste, on its own and out of context, is interpreted as a sign of white wealth only by the excessively naive; like the Silicon Valley hoodie uniform, it is a contextual move which is able to signify only against an understood landscape of signification; it can *only* be considered in context.[2]

Much related, a cargocult is the confusion of surface details and instantiation-specific components for substantive or functionally necessary aspects (Feynman 1974, Reason 2016). “The cargoculter builds a motorless airplane from palm fronds, sprinkles it with holy water, and prays to the gods for it to fly” (Reason 2016). Typically the confusion is born of a lack of deep systems understanding of the target domain. As a result, a cargocult imitates superficial and aesthetic elements (markers or ritual indicators) in expectation of their efforts reproducing the operation of the original. These confusions may be attributed in part to confusion over the direction of causality, and the role of the components in the enveloping system. A system which surrogates non-causal attributes, and especially the surface products of deeper causes, can be considered a cargocult.[3]

Kahneman and Tversky have also theorized *attribution substitution*, in which an agent tasked with a difficult question may resort—unwittingly—to answering a related, proxying but distinct question that is easier to answer. Perhaps most famous as an example is the bat and ball cognitive reflection test, where subjects are asked to calculate the cost of a ball, given that the bat and ball together cost $1.10, and the bat costs $1 more than the ball: respondents appear to most immediately answer that the ball costs ten cents, which the behavioral economists speculate is the result of subjects substituting the real task for the task of merely parsing large and small quantities (e.g. as a fast-and-frugal heuristic for making financial decisions). Whether subjects come to the correct response is largely reliant on whether they use System 2 thinking to monitor and correct their System 1 intuition. While analogous in underlying mechanism to *surrogation*, attribution substitution in Kahneman’s factoring is performed quickly and unconsciously, rather than at the level of conscious institutional or individual structuring. Still, we can think of them as similar in kind.

Finally, in artificial intelligence research (which frequently mobilizes Goodhart's Law), there are the concepts of *underspecification* and *nearest unblocked strategy* (Manheim 2019, Arbital 2020). Specifying a telos in code proves a hard task: any behaviors not explicitly prohibited may be exploited; incentives turn perverse; roadblocks prove insufficient and—like Midas—the goal literal turns out not to be the goal actual.

In his December 2018 article on the origins of Goodhart's and Campbell's laws, Jeff Rodamar makes the case that differences in use of the different terms, field to field, "harms communication, creates barriers to science, and hinders improvements in practice." This is to say nothing of the many other similar concepts, enumerated above, in management accounting, behavioral economics, artificial intelligence, and the philosophy of games.

2. Behavioral vs. introjective surrogation

Roughly, there are two kinds, or possible stages, of surrogation: the alteration of behavior towards a surrogate, and/or the psychological internalization of the surrogate—the reification of the surrogate *as if* it were the thing itself; an amnesia surrounding the switch.

Perhaps the major problem of surrogation is that it alters and corrupts human behaviors, moving their telos away from the originally desired behavior and into those which, while rational at an individual level (as exploitations of the incentive structure), are inefficient at the level of institution and society. Moreover, in cases in which individuals are aware that their incentivized behavior diverges from pro-social goals, the activity loses meaning and the individuals become "cynical"—they are aware of the performative aspects of their acitivity. (See, by way of example, Michael Inzlicht's reflections on his disillusionment with social psychology [2016].) In selectively rewarding individuals or institutions who optimize *away* from the real target and toward the instituted surrogate, it (1) discourages play interested in the real target, (2) discourages players interested in the real target, who engage in exit from the game, (3) increasingly promotes and advances individuals or institutions who optimize (undesirably) to the surrogate at the cost of the real target.

To understand just how *perverse* (*cf.* "perverse incentives") surrogate incentives can be, we can look to Robert Jackall's sociological study of institutional ethics, *Moral Mazes*:

> [A]t Covenant Corporation the story is told about a plant that produced a useful by-product at no extra cost. One simply had to store it until it was needed for other internal operations. Covenant, however, works with an accounting system that considers by-products as inventory; moreover, inventory counts against one at the end of a fiscal year. In order to cut costs,

> managers decided to throw out the by-product at the end of a financial cycle. But a sudden shortage of the material trebled its cost two months later. To service their own operations, managers had to go hat in hand to their competitors to buy the material at the premium prices.

The introjective kind of surrogation is performed (or "happens to") not just those are expected themselves to optimize toward the surrogate—as in the behavioral kind—but also those who have constructed the incentive structure, who pass down the surrogate and, in institutional contexts, may even have designed it. In more decentralized social settings, subject to cultural inheritance and the ongoing, distributed negotiation of norms—realms where we "are organized beings, but are not the authors of our organization" (paraphrasing Noë 2015)—values, preferences, and norms are more amorphous, enacted but often only half-consciously known; their slow replacement by a surrogate can occur without intentionality or conscious recognition.

Goodhart and Campbell's Law approach surrogation *from the perspective of the system*. Nguyen's intervention, with his concept of "value capture," is to shift the perspective away from the system and toward the individual inhabiting it.

3. Quantitative surrogation: The problem of value capture

By setting a metric as a target, and by linking that target to a reward structure, we create an incentive for that metric to be gamed in some way. (Rodamar 2018)

The central problem in any superorganism or institution is that of aligning values between members so as to coordinate action toward a shared purpose; this dilemma is known as the principal-agent problem. Solving principal-agent problems requires preferential treatment—rewards or punishments doled out on the basis of performance (fixed-rate salary with bonuses is a classic example of financial incentive—though prestige and reputational incentives have proved efficacious on their own).

Broadly speaking there are three main advantages to instituting measurement across these systems, David Manheim writes in his essay series on measurement for *Ribbonfarm*: "[It] replaces intuition, which is often fallible. It replaces trust, which is often misplaced. [And it] finesses complexity, which is frequently irreducible"—where *irreducible* entails *intractable*[4] (2016a, 2016b). In other words, the organization wishes to *supervise* and *monitor* the behavior of its subagents, to ensure honest and high quality performance. Complexity must somehow be reduced to a synopsis, or indicator, in order to effectively evaluate performance. Additionally, a desire to make the basis of this preference consistent across

the organization, and transparent and legible for involved parties—as is frequently expected in a society that values equal opportunity—involves instituting public bases for advancement. In some arenas, performance plays out in a way that is easily quantitatively tractable, such as sales figures, but where there is divergence—where a number or statistic is preferred over "the real deal"—the publicity makes the surrogate gameable, and rational self-interest adjusts its targets accordingly. Our accounting of the motivations for quantitative surrogation thus includes not only the replacement of intuition and the reduction of complexity, but the production of legible, transparent, consistent, fair, and objective-seeming bases to ensure better management of the organization, and (contiguously) the principal-agent solving preferential treatment of organization members. Nguyen, in *Games and the Art of Agency*, illustrates how the reduction of qualitative values to quantitative metrics ensures the "units" or demonitator of evaluation are consistent across both time and space:

> large-scale institutions often need quantified measures of their various functionings for management purposes. High-level administrators in large institutions need to be able to compare, say, productivity, customer satisfaction, and worker satisfaction across various departments. This requires quantified representations of values. An administrator might first need to aggregate productivity numbers across different departments in, say, their Tokyo and Los Angeles locations, or aggregate productivity numbers from all locations to compare institutional productivity over years.

3.1. Complexity and compression

What this involves, necessarily, is the reduction or summarization of complex and fuzzy realities.[5] Manheim:

> Complex systems have complex problems that need to be solved. Measures can summarize, but they don't reduce the complexity. This means that measures hide problems, or create them, instead of solving them . This concept is related to imposed legibility, but we need to clarify how in a bit more detail than the 'recipe for failure' discussed in the linked piece. In place of that recipe, I suggest another triad to explain how complexity is hidden and legibility is imposed by metrics, leading to Goodhart's law failures. These failures are especially probable when dimensionality is reduced, causation is not clarified, and the reification of metrics into goals promotes misunderstanding. (2016a)

In the language of computation, the surrogate *lossily compresses* some complex whole, and bears inverse fidelity to its divergence from said whole, and conflates many possible worlds into a single measure. For instance, one's appearance as a student with a GPA of 3.15 may mask two very different realities—on the one hand, a straight-B student; on the other hand,

an enormously talented physicist who flunked his compulsory literature course. The Australian counterinsurgency expert David Kilcullen writes in "Measuring Progress in Afghanistan":

> Violence tends to be high in contested areas and low in government-controlled areas. But it is also low in enemy-controlled areas, so that a low level of violence indicates that someone is fully in control of a district but does not tell us who.

When quantitative metrics obscure meaningful valence differences in the compressed whole, selection based on these metrics produces disastrous, counterproductive results which are difficult to monitor precisely *because* the disaster is invisible to the system of monitoring. Historian Muller, in *The Tyranny of Metrics*, describes the downfall of simple counting:

> From [commanding officers'] point of view—and from the point of view of the politicians to whom they reported—every arrest was of the same value. The course of action that produced the best performance indicators did little to diminish the sale of narcotics. (2018)

We can adapt the excerpt for psychology—whose use of surrogation, as we will soon see, has led to its generalizability and replication crises:

> From the point of view of researchers' hiring committees and grant foundations, every publication was of the same value after controlling for the prestige of the journal. The course of action that produced the best performance indicators did little to advance the discipline's larger project of quality research.

3.2. Behavioral surrogation: Motivated interpretation & massaged maps

While quantitative surrogates are often instituted to provide an objective oversight (to prevent being fooled by an employee's "spun" self-representation, for instance), in practice, when the interpretation of reality *as* statistics—the choice of how to compress that reality—is left to the monitored agents, statistics are easily "massaged." Muller reports:

> In 2014, a whistle-blower from the London police force told a parliamentary committee that massaging statistics had become "an ingrained part of policing culture": serious crimes such as robbery were downgraded to "theft snatch," and rapes were often underreported so as to hit performance targets. As a retired detective chief superintendent put it, "When targets are set by offices such as the Mayor's Office for Policing and Crime, what they think they are asking for are 20% fewer victims. That translates into 'record 20% fewer crimes' as far as… senior officers are concerned." Such underreporting and downgrading of crimes "are common knowledge at every level

> in every police force within England and Wales," he added. (2018)

Muller quotes a Chicago detective on the ease of "juking the stats" (i.e., orienting "the activity of the department toward seemingly impressive outcomes"):

> "It's so easy [to massage figures]." First, the responding officer can intentionally misclassify a case or alter the narrative to record a lesser charge. A house break-in becomes "trespassing"; a garage break-in becomes "criminal damage to property"; a theft becomes "lost property."

This massaging occurs at all levels, so long as the information is being passed upward in command, i.e. to the agents responsible for doling out preferential treatment. Since in many cases even the highest-ranking members of an organization are responsible to shareholders or a public, massaging occurs at all levels. And while superiors often prefer accurate over massaged information in order to make better strategic decisions and fill out the organization with competent workers, in some situations, statistics massaged at lower levels may be preferable or even knowingly demanded, since "keeping their hands clean" in this way allows higher-ups plausible deniability in passing their own claims forward.

3.3. Introjective surrogation: Value capture, value clarity

Games extract pleasure from what C. Thi Nguyen, in 2020's *Games and the Art of Agency*, calls "value clarity":

> Life is a confusing welter of subtle values, in a vast and confusing plurality. Living our lives, as fully sensitive valuing agents, involves making painful judgments, tough decision calls, and agonizing comparisons.
>
> In game life, our temporary agency's values are usually extremely clear. That clarity is encoded into a game's specification of its goals. The values we take on in games are clearer, easier to apply, and easier to evaluate than our enduring values.

Game play, in other words, involves an "all-consumingly instrumental mode of practical reasoning." The legibility, meanwhile, allows public ranking, encourages improvements in productivity and performance by establishing common knowledge of relative performance, fostering competition among members.

The appeal of value clarity can lead human superorganisms into what Nguyen calls *accidental gamification*, where game-like features—such as

clear metrics, often introduced top-down with the explicit aim of motivating employees through public competition:

> [A]cademic life has recently come to be ruled by quantified metrics for research quality—like citation rates and impact factors. These metrics may not have explicitly been designed to produce gamification among researchers. Conceivably, they arose from the bureaucratic need to collate information, or in university administrators' quest to make more object-sounding decisions about faculty hiring and promotion. But the clear, simple, and quantified nature of such metrics can foster game-like motivation… We could be drawn to redefine our notion of success in the newly clear terms specified by those metrics. (2020)

The gamification of academia, science, and the "global knowledge game" is discussed in the following section, "Surrogation and the crisis in psychology."

> Value capture occurs when:
>
> 1. Our values are, at first, rich and subtle.
> 2. We encounter simplified (often quantified) versions of those values.
> 3. Those simplified versions take the place of our richer values in our reasoning and motivation.
> 4. Our lives get worse.

In "simplifying the specification of the target" we end up pursuing, "with ever more ferver and ferocity, the wrong target." Often, by the laws of complexity—that is, the inevitability of perverse incentives—surrogated efforts even make the situation worse than passivity—this being part of the case made by Michael Huemer in his defenses of policy passivity (2012).

> by simplifying the specification of the target, we may bring outselves to pursue, with ever more ferver and ferocity, the wrong target.
>
> Such measures are useful, but we must always recall that they are merely abbreviations—usefully portable simplifications of something larger and subtler. But when our values are captured, we are motivationally caught by a simplified measure.

4. Surrogation and the crisis in psychology

As we have seen, surrogation permeates distributed human projects, or "superorganisms"—institutions like the military, police department; the medical or justice system; diplomacy and trade; but also what Sarah Perry

(2020) dubs the “global knowledge game”—the ongoing process of attempting discovery of global truths spanning scientific and, to a lesser extent, humanities work. In other words, a lifting of knowledge *out of context* and into some human (aspirational-)universal or generalization.

There are numerous surrogation-caused problems in the global knowledge game (GKG). Because the GKG has become a vast enterprise characterized by information overload—by the simultaneous production of millions of members—and because there is a vast, distributed incentive structure designed to reward certain behaviors ostensibly in the service of knowledge production, we should expect it to have the same institutional issues of stats-gaming (e.g. p-hacking) already discussed with respect to the police and military. (Moreover, surrogation is common across knowledge-oriented fields, such as education, where in the United States we’ve seen controversies over “teaching to the test” as well as more blatantly corrupt Goodhartian actions such as teachers manually altering students’ Scantron forms.)

Additionally, in the “inexact sciences”—that is, those which are attempting to mature past their qualitative roots and into a more quantitative or empirical science, for instance psychology’s abandonment of phenomenology and psychoanalysis in favor of statistical lab studies—there is a problem of wanting to grow up too fast. In their rush to “objectify” and rigorize themselves, many of the social sciences have hastily abandoned old methods, replacing them entirely with a more performatively “scientific” surrogate. Here, I’ll use Tal Yarkoni’s recent assault on social psychology, “The generalizability crisis” (2019), as a launching pad to discuss the phenomenology or psychology *of* surrogation, as well as some of the sociological reasons that institutions deep in surrogated divergence (i.e., away from the “real” target) are so difficult to correct.

4.1. Surrogation and the replication crisis

[TK: Gigerenzer on the “surrogate idol” of a universal method; p-hacking and gamification of paper submission; the incentive structure that discouraged replication in the first place.]

4.2. Surrogation and the generalizability crisis

The broad argument Yarkoni advances is that psychology studies’ ability to generalize—for the narrow bounds of a lab study done with “just one video, one target face, and one set of foils” to provide evidence for the existence of some broad psychological construct like ego depletion—is orders of magnitudes lower than traditionally assumed in the field. Yarkoni’s critiques are not new—as he himself notes, many thinkers in the inexact sciences have been raising the alarm on similar issues, including Gerd Gigerenzer and Paul Meehry, in some cases for upwards of half a century—but they compile and make sense of the scope of the problem social psychology faces.

First, a psychological construct, in order to gather evidence as to its "existence" or "nonexistence"—and even here there is a whiff of conceptual confusion—must be operationalized:

> things like cognitive dissonance, language acquisition, and working memory capacity—cannot be directly measured with an acceptable level of objectivity and preccision. What *can* be measured objectively and precisely are operationalizations of those constructs—for example, a performance score on a particular digit span task, or the number of English words an infant has learned by age 3. Trading vague verbal assertions for concrete measures and manipulations is what enables resaerchers to draw precise, objective quantitative inferences; however, the same move also introduces new points of potential failure, because the validity of the original verbal assertion now depends not only on what happens to be true about the world itself, but also on the degree to which the chosen proxy measures successfully capture the constructs of interest—what psychometricians term construct validity.

Yarkoni himself has characterized the surrogative aspects of operationalization: the validity of any findings depend, post-operationalization, on "the degree to which the chosen proxy measures successfully capture the constructs of interest."

Once the study is completed, a second stage follows: the discovered quantitative or operationalized reality is re-translated back into language via generalization or loose induction. The coarse metrics to some extent "disappear," as we re-enter the realm of language where knowledge is hosted and decisions made. The context is further stripped as the narrow lab finding is "generalized" into a larger claim about human behavior: "Papers should be given titles like 'Transient manipulation of self-reported anger influences small hypothetical charitable donations,' and not ones like 'Hot head, warm heart: Anger increases economic charity.'," Yarkoni writes.

4.2.1. Appearance-optimization as a cargocult

The important thing, it appears, is that the numbers have the right form. (Yarkoni 2019)

Recall that to *cargocult* is to imitate a work's surface structures while lacking a proper understanding of the actual mechanisms behind its power. This kind of behavior can be either opportunistic and knowing, putting on a show of appearances for others—as in the cult leader, cynic, or grifter—or else merely a kind of magical thinking and wish fulfillment: "The cargoculter builds a motorless airplane from palm fronds, sprinkles it with holy water, and prays to the gods for it to fly." The psychologist builds up all the meticulous appearances of real science, and prays that his findings contribute to human knowledge. What's more, since we live in a society that

unwittingly or uncaringly surrogates appearance for reality in decision-making and evaluation—in other words, an *optikratic* society[6] that lives and dies by appearances—these performances frequently *do* succeed in "flying," perpetuating the optikratic incentive structure.

Yarkoni himself uses the phrase "cargocult science" to refer to the performative aspects of empiricism in psychology, and its concurrent optimization of metrics à la p-hacking:

> It's hard to think of a better name for this kind of behavior than what Feynman famously dubbed *cargo cult science* (Feynman, 1974)—an obsessive concern with the superficial form of a scientific activity rather than its substantive empirical and logical content.

Here, the "superficial" stands as the actually-incentivized surrogate, and the "substantive" the surrogated destination which organizations and players in the global knowledge game self-purport to navigate toward.

Ironically, it may be the case that the inexact sciences, rather than abandoning qualitative research, have merely cloaked it in the grand rhetoric of empiricism; Yarkoni concludes that "many fields of psychology currently operate under a kind of collective self-deception, using a thin sheen of quantitative rigor to mask inferences that remain, at their core, almost entirely qualitative."

4.2.2. The sociology of surrogation: Why can't honest actors recover corrupted systems?

Researchers, Yarkoni writes, are driven in psychology and related fields "to expend enormous resources on studies that are likely to have very little informational value even in cases where results can be consistently replicated." Statistically and inferentially unfounded claims are passed up, from psychology, to the highest levels of public and private decision-making, altering the behavior of governments, corporations, and public institutions alike, in large part because this performance of empiricism is highly effective in lending legitimacy to psychological hypotheses. Books are published, and become bestsellers, or talks given that go viral, by psychologers who lead the public to claims and generalities that their studies do not support. There is widespread abuse and gamification of statistics of legitimization, the most well-known being p-hacking. Yarkoni presents a number of "next steps," given this horrifying state of affairs, but they are designed for individuals: leave the field, practice slower science, present one's findings more modestly. As a result, they miss the crucial sociological angle from whence these problems originate. There are game-theoretic forces at play here, and the structure of incentives in which the problematic behavior originates is not much altered by individual decision-making.[7]

The first problem is that more modest claims come at the loss of power, prestige, and reputation. Not only would the field be ceding much of its previously claimed credibility, but that credibility would ostensibly drop even further on the basis of the prior deception. It would arguably take quite some time for the field's place in public discourse to recover.

The second problem is that as individual psychologists leave the field—and this is already happening, at least insofar as graduate students high in integrity are turned off from psychology's performative pseudoempiricism—as individual psychologists leave the field, or cease to advise public policy, or cease to make grand claims on-stage, they will be replaced by those willing to. Those who replace will on average be those with less integrity, less interest in rigorous skepticism, and less knowledge as to the limitations of their practice than those who they replace. They will then train PhD students in their techniques.

In other words, as knowledgeable insiders slowly leave the field (or choose never to join it in the first place), psychology will become increasingly dangerous and destructive until its public credibility collapses entirely. This process has been with the discipline from the beginning; academic psychologists Yoel Inbar and Michael Inzlicht report multiple occasions of "bright undergraduates" voicing complaints similar to Yarkoni's, and we can only imagine that psychology's inability to convincingly answer such concerns discourages those with the foresight to see it from entering. In other words, we have both a selection problem and a self-selection problem.

Social psychologist Pamela Smith; interview on *Two Psychologists Four Beers* (recall Bourdieu's idea that the gossip of a field makes up some of its essential wisdom):

> We are still rewarding people based on publications. It is true now more than ever, if you want a publication, maybe people are paying more attention to sample, but they're very happy to let people do online studies that don't necessarily map well onto behavior, and just run a bunch of them. You get penalized if you want to do careful work. You get penalized if you want to do work on people other than college undergraduates or people who are willing to do online surveys for fifty cents a shot.

A third problem, related to the second, is that those psychologists who choose to stay will be out-competed, out-hired, and out-tenured compared to those who are willing to play ball with p-hacking regimes, with performative pseudoempiricism, and with the publish-or-perish emphasis on quantity over quality. Misuse raises the bar of expectation; those who optimize toward "real" science—in other words, the surrogated target—are penalized in their competition with those who more efficiently and directly optimize toward the actual metrics of promotion, advancement, and recognition—the surrogate that is "optics." This incentive structure is real and affects not just the career prospect of individuals but the larger efficacy and service of the institution.

Finally, psychology—insofar as it can be meaningfully said to "freeride" the reputation of legitimate science, by enjoying the benefits of its perceived reputation while showing little obligation to the same rigor—will increasingly harm the overall perception of legitimacy of the sciences. We can see some taste of this in the so-called Science Wars of the late 20th C, where the failings and hubris of social sciences helped delegitimize the "hard" science, in part because the problem of psychology is the problem of inference, while of a very different *scale* of problematicity compared to physics, are of the same kind—the seemingly intractable problem of inference.

4.2.3. Surrogation vs. contextualization

There is good reason that Yarkoni and Paul Meehl both emphasize that much of the current crisis in psychology comes from the conventional, automatic, and uncritical surrogation of statistical measures. The alternative to surrogating the qualitative-holistic is contextualizing the metric *within* the qualitative-holistic, using the two as mutual agitation, a dialectic.

Muller 2018:

> Interpretation of indicators is critically important, and requires informed expert judgment. It is not enough merely to count incidents or conduct quantitative or statistical analysis—interpretation is a qualitative activity based on familiarity with the environment, and it needs to be conducted by experienced personnel who have worked in that environment for long enough to detect trends by comparison with previous conditions.

5. Qualitative surrogation: The tragedy of appearances

Though surrogation has been usually described in terms of the lossy compression from qualitative, inuitive, holistic judgment, and into quantitative metrics, it can occur any time a marker is substituted for what it demarcates. In signaling theory, classically, signals are external, public-facing attributes that indicate, to other organisms, a probabilistic presence of some hidden, private trait. Just like in language, with the connection between the *signified* and the *signifier*, this ability to "stand proxy for," and represent publically, some private and hard-to-verify truth is built up through brute associative learning: an experience with the coincidence of some prominent physical marker and some attribute instill a relationship that can be meaningfully used as the basis for future inference. To illustrate just how common this kind of behavior is in our social lives, consider: how do we size up an strangers's socioeconomic class, or extrapolate a candidate's future performance in a hiring interview?[8]

Unfortunately, the metonymic surrogation and reification we see play out in the sphere of metrics plays out in the qualitative signaling sphere as well. To take an example from the history of pop music, authenticity—a hard-to-measure, complex trait—has seen itself instantiated in different ways, for instance, the folk scene in Greenwich Village in the 1950s was perceived as having this reputation; the same is true in the late 20th and early 21st century of "lo-fi aesthetics"—music recorded on relatively inexpensive amateur equipment. The logic for this association was relatively straightforward, if not premised on costly signals but rather, the lack of incentives present in these domains—folk singers typically were single individuals, making almost no money, requiring only a guitar and a small performane venue (e.g. a bar or comedy club); musicians home-recording from Tascam 4-Trax did not need to pay a studio or producer's fee, which means not needing label support. In both cases, there is a lack of financial pressure, with the recognition that such pressure tends to corrode or compromise an audience ideal of "aesthetic integrity"—the vision of the artist, rather than a catering to the listener.

When such fields of production were ignored, and there was no money available for their agents, there was a meaningful sense in which these associations were costly: artists which cared more about autonomy would forego the income and reputation that label support might afford them. When the scenes began to attract attention, however, there was a quick free-rider effect of *acting as if*: there was nothing intrinsic to performing on an acoustic guitar, or having audio distortion due to poor compression capabilities of recording hardware, that was more "honest," and by imitating all the aesthetic residue and markers—the associated surface signals—of authenticity, acts would see authenticity conferred on them in turn.

This burgeoning fetishization of surface aesthetics still permeates the independent music scene, where tape warble and white noise, vocal clipping and compression, is deployed tactically to give a certain affective impression—and since the affect is so fleeting, who could make an accusation of falsehood "stick"? This is one case of surrogation: by incentivizing compliance to a set of surface qualities, in a purported bid for monitoring and securing authenticity, musicians and labels are, in actuality, ironically encouraged to falsify their own material origins and capacities.

It was against this backdrop we can understand Dylan's 1965 performance at Newport Folk Festival—an incident with its own encyclopedia page, the "Electric Dylan controversy," and the flipside to this surrogation. We can see footage today of the set: Dylan, performing the exact same songs that had been heralded, and borderline sanctified, for their honesty and activism, but performing with an electric, rather than acoustic guitar. Dylan had "plugged in"; the widespread sentiment was that in doing so, Dylan had "sold out," was no longer a performer of integrity, on the basis of a new guitar sound. Without playing down the complexities of the historical situation—without denying that there is something legitimate about anger over symbols, and that the mythologization of this event undoubtedly has led it to be exaggerated—how else can we make sense of the outrage that followed, than as the reification of an associated but causally distinct

measure, than as the surrogation of a complex trait like “authenticity” for a much simpler one, the way one speaks or the instrument one plays? The reception lasted for years in Dylan’s tours, jeers of “Judas” from the crowd.

The imitation of surface attributes, rather than causal mechanisms, is a common one in beginning artists. In Arthur Danto’s book-length profile of Warhol, we encounter Warhol’s early imitation of AbEx “paint drips,” his belief that it was somehow critical to the painting project:

> [He] applies paint the way an Abstract Expressionist artist would, allowing it to drip. “You can’t do a painting without a drip,” he told Ivan Karp, who was director of the Castelli Gallery. This is what I meant by saying that he used Abstract Expressionist gestural painting as protective coloration. The drips did not come from some inner conviction… (or, we might interpret, an internal logic) …they did not refer to that moment of trance when the A. E. painter moved the paint around without tidying up. “The drip”… for Warhol… [was] an affectation…

Whereas, for the original Abstract Expressionists, paint drips were a byproduct of a *technique* that embodied an *ideology* of art (an ideology much in line with the emphasis on spontaneity and honesty found also in folk music). But here that very byproduct is lifted out of its context and treated as a goal in its own right. Amidst these performances, which are often enough to fool critics, genuine embodiments of qualities like innovation or integrity go unrecognized, while regurgitation disguised by savvy signaling is showered in praise. Today in many visual art cultures, the aesthetics of a “zine”—themselves artifacts of copymachine technologies from the 1990s, as pioneered by groups like Riot grrrl—surrogate the proxied-for qualities, and are perceived as somehow “more DIY” than those projects made with contemporary projects. Filmmakers who wish to be perceived as experimental will engage in the now-antiquated techniques of avant-gardes past, in order to seem “of a kind” with their hallowed paters.

This historical residue is all around us—it is the lingering ooze of prestige past, available for any who care more about said prestige than the field. We can call its effects *retrolegitimation*. And yet, considered this way—as the anemic surrogate, a pretense *as if*—the appeal to retrolegitimation, and the presence of this residue in works ought to be treated as a reverse indicator, as zombie art animated by the hungover associations of eras past. AD Jameson describes the dynamic:

> The canonical works define the style and range of [what is considered “proper” U.S. experimental] cinema: It is non-narrative (favoring surreal logic or structural organizing principles), abstract, often incorporates found footage, and also frequently involves directly treating the film itself (scratching it, painting it, growing mold on it, and so on). It often

> demonstrates some aspect of the film apparatus or filmmaking process, sometimes by taking a self-reflexive approach (foregrounding the use of the camera) or a conceptual approach (projecting through alternate substances, or projecting plain black leader, or projecting nothing but the projector light itself).

Imitation of a canon is obviously antithetical to the spirit of experimentalism. And yet "the film students of today frequently make work that employs those techniques [associated with historical experimentalism]. The question then becomes: Are they making experimental films?" We can leave quibbling over labels to art historians while confidently assessing that the original target of experimental practice has been lost, surrogated for those techniques which are known, in the critical and public sphere, to have accompanied it—and which are still met, by critics and elite audiences, with the prestige accorded the originals.

And is against this backdrop—the nefarious surrogation of real efforts into cardboard cutouts, surface signaling replacing genuine embodiment—that we can understand the emergence of showy, fantasy-ridden, egoic and artificial glam rock in the early 1970s, as well as the disdain that it raised. The pop studies scholar Simon Reynolds, in his book on glam *Shock & Awe*, sets the scene for us with an illustration of surrogation in 60s theater:

> a post-Method school of actors and directors aspired to a de-theatricalised form of naturalistic acting, all mumbling and tics, that inevitably spawned a new set of mannerisms that today look as stagey and trapped in time as the Hollywood golden age of poise and elocution. In all the arts, in fact, every attempt of realism, no matter how stringently stripped down or crude, seems to birth a new repertoire of stylised conventions and stock gestures. Bowie, for one, was acutely aware of this in relation to rock, which he precociously grapsed was a *performance* of real-ness rather than a straightforward presentation of reality onstage.

This is both in the sense that all naturalness is "technically" a performance, and also that the performance had become increasingly and meaningfully more conscious, strategic, and commercial. Glam, as Reynolds shows, took the strongest symbols of Sixties "natural honesty"—hair and nudity—and mocked them with makeup, costume, and dye. What it was really mocking was surrogation—the dangerously cheerful illusion that we can fetishize a measure and it will continue to tell us the truth about its subject. How else can we understand these great developments in the history of pop, other than as products of freeriding and surrogation, of symbols reified as the things themselves?

6. Mitigations

6.1. Supplementation vs. surrogation

Measurements alone cannot devastate; they are merely a second source of information. It is when other forms of evaluation are surrogated to measurements, and become instituted metrics, that causes trouble—we now have less information rather than more.

Manheim (2016a):

> On the one hand, Kahneman found that decisions are subject to cognitive biases and can be systematically improved once we move past our intuition. On the other, despite our systematic biases, as Gary Klein originally noted when studying firefighters, many decisions don't use metrics, and are incredibly effective despite that. In fact, this success isn't despite the lack of cognition, but because of it. Klein's "recognition-primed decision making" works exactly where our intuition beats measurement. As Klein and Kahneman now agree, there are domains in which "raw intuition" beats reflection.
>
> [Hubbard] opens the book saying that "no matter how 'fuzzy' the measurement is, it's still a measurement if it tells you more than you knew before.

The issue isn't the introduction of metrics, which have in fact gotten rid of many human biases, the personal grudges of managers, the inconsistencies of distributed bureaucracies, etc. It's the substitution of holistic qualitative eval for reductive quantitative eval.

But treat them as complementary, you get somewhere. Metrics strip context, but allow the evaluated employee to re-insert context by narrativizing their metrics? We're starting to iron out the kinks.

Similarly, employees self-narrativizing is great for qualitative richness, but shitty when it comes to reliable, grounded, no-bullshit understandings of performance. Metrics ground them in reality, and shrink the space of fabricated unrealities.

We should've expected this in the first place, right? It's basic Pareto frontier stuff. Single-minded solutions have serious drawbacks, no matter which you pick. Start combining them and you start getting balanced feedback.

Yarkoni:

> It "is the mismatch between our generalization intention and the model specification"—in other words, between holism and quantitative representative—"that introduces an inflated risk of inferrential error, and not the model specification alone." (Yarkoni 2019).

6.1.1. System 1 and System 2

6.1.2. Developing an aesthetic

Quoting Ben Connable's Rand study *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*:

> It would be difficult (if not impossible) to develop a practical, centralized model for COIN assessment because complex COIN environments cannot be clearly interpreted through a centralized process that removes data from their salient local context.
>
> Connable characterizes counterinsurgency as "both art and science, but mostly art." […] The tendency is to treat as pure, measurable science what is of necessity largely a matter of art, requiring judgment based on experience.

6.2. Tracking intuitions

An alternative approach to combining System 1 and System 2 thought—and a proven effective approach to measurement—is to track locals' subjective, acted-upon, skin-in-the-game assessments of complex situations, e.g. tracking the stability of Afghanistan through the price of market goods:

> Afghanistan is an agricultural economy, and crop diversity varies markedly across the country. Given the free-market economics of agricultural production in Afghanistan, risk and cost factors—the opportunity cost of growing a crop, the risk of transporting it across insecure roads, the risk of selling it at market and of transporting money home again—tend to be automatically priced in to the cost of fruits and vegetables. Thus, fluctuations in overall market prices may be a surrogate metric for general popular confidence and perceived security. In particular, exotic vegetables—those grown outside a particular district that have to be transported further at greater risk in order to be sold in that district—can be a useful telltale marker.

(This is similar to the approach in ethnomethodology, which attempts to understand sociality through demonstrated behavioral adaptation and response to stimuli.)

6.3. Minimizing measurement surrogation (Goodhart-Campbell)

Here we can understand *G* as the surrogated original and *G** as the surrogate.

From *LessWrong* (2010):

Hansonian Cynicism

Pointing out what most people would have in mind as *G* and showing that institutions all around are not following *G*, but their own convoluted *G*s*. Hansonian cynicism is definitely the second step to mitigation in many many cases (Knowing about Goodhart's law is the first). Most people expect universities to be about education and hospitals to be about health. Pointing out that they aren't doing what they are supposed to be doing creates a huge cognitive dissonance in the thinking person.

Better measures

Balanced scorecards

Taking multiple factors into consideration, trying to make *G** as strong and spoof-proof as possible. The Scorecard approach is mathematically, the simplest solution that strikes a mind when confronted with Goodhart's law.

Optimization around the constraint

There are no generic solutions to bridging the gap between *G* and *G**, but the body of knowledge of theory of constraints is a very good starting point for formulating better measures for corporates.

Extrapolated Volition

CEV tries to mitigate Goodhart's law in a better way than mechanical measures by trying to create a complete map of human morality. If *G* is defined fully, there is no need for a *G**. CEV tries to do it for all humanity, but as an example, individual extrapolated volition should be enough. The attempt is incomplete as of now, but it is promising.

Solutions centred around Human discretion

Human discretion is the one thing that can presently beat Goodhart's law because the constant checking and rechecking that *G* and *G** match. […] However, this is not scalable in a strict sense because of the added testing and quality control requirements.

Left Anarchist ideas

Left anarchist ideas about small firms and workgroups are based on the fact that hierarchy will inevitably introduce goodhart's law related problems and thus the best groups are small ones doing simple things.

Hierarchical rule

> On the other end of the political spectrum, Molbuggian hierarchical rule completely eliminates the mechanical aspects of the law. There is no letter of the law, its all spirit. I am supposed to take total care of my slaves and have total obedience to my master. The scalability is ensured through hierarchy.

7. Works Cited

Amodei and Clark (2016). "Fauly Reward Functions in the Wild." *Open AI Blog*. https://openai.com/blog/faulty-reward-functions/.

Arbital (2020). "Nearest unblocked strategy." https://arbital.com/p/nearest_unblocked/.

blogospheroid (2010). "The Importance of Goodhart's Law." *AI Alignment Forum*. https://www.alignmentforum.org/posts/YtvZxRpZjcFNwJecS/the-importance-of-goodhart-s-law.

Choi, Hecht, and Tayler (2012). "Strategy Selection, Surrogation, and Strategic Performance Measurement Systems". *Journal of Accounting Research*. 51 (1): 105–133. https://doi:10.1111/j.1475-679X.2012.00465.x.

Choi, Hecht, and Tayler (2011). "Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation." *The Accounting Review*. 87 (4): 1135–1163. https://doi:10.2308/accr-10273.

Feynman, Richard (1974). Caltech commencement address.

Huemer, Michael (2012). "In Praise of Passivity." *Studia Humana* 1,2 (2012): 12-28.

Inzlicht, Michael (2016). "Reckoning with the past." *Getting Better*. http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past.

Jackall, Robert (1988). *Moral Mazes: The World of Corporate Managers*. Oxford University Press.

Kahneman and Tversky (1973). "On the Psychology of Prediction". *Psychological Review*. 80 (4): 237–51. https://doi:10.1037/h0034747.

Manheim, David (2016a). "Goodhart's Law and Why Measurement is Hard." *Ribbonfarm*. https://www.ribbonfarm.com/2016/06/09/goodharts-law-and-why-measurement-is-hard/.

Manheim, David (2016b). "Overpowered Metrics Eat Underspecified Goals." *Ribbonfarm*. https://www.ribbonfarm.com/2016/09/29/soft-bias-of-underspecified-goals/.

Manheim, David (2019). “Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence.” *Artificial Superintelligence: Coordination & Strategy.*

Perry, Sarah (2020). “Ignorance: a skilled practice.” *Carcinisation*. https://carcinisation.com/2020/01/27/ignorance-a-skilled-practice/.

Nguyen, C. Thi (2020). *Games And The Art Of Agency*. Oxford University Press.

Noë, Alva (2015). *Strange Tools: Art and Human Nature*. Hill and Wang.

Ortega and Maini et al (2018). “Building safe artificial intelligence: specification, robustness, and assurance.” *DeepMind Safety Research*.

Reason, Suspended (2016). “Intro to Cargocult.” *Suspended Reason*. https://suspendedreason.com/2016/11/16/intro-to-cargocult/.

Reason, Suspended (2020). “The Dark Miracle of Optics.” *Less Wrong*. https://www.lesswrong.com/posts/zzt448rSfwdydinbZ/the-dark-miracle-of-optics.

Rodamar, J. (2018). “There ought to be a law! Campbell versus Goodhart.” *Significance*, 15(6), 9–9. https://doi.org/10.1111/j.1740-9713.2018.01205.x.

Simler, Kevin (2016). “Minimum Viable Superorganism.” *Ribbonfarm*. https://www.ribbonfarm.com/2016/02/11/minimum-viable-superorganism/.

Yarkoni, Tal (2019). “The Generalizability Crisis.” PsyArXiv. November 22. https://doi:10.31234/osf.io/jqw35.

8. Footnotes

---

1. In this way, the realm of social reputation is prone to a range of related dynamics in which the gap between appearance—assumed signals or markers of some hidden reality—and that reality itself causes problems. ↩

2. Partially this is described by the “barberpole” metaphor of fashion, “where lower classes continually imitate higher classes, who are themselves engaged in a continual quest for “distinction” from the chasing masses… Its cyclical nature is the result of limited options and a continual evasion of freeriders who exploit an associative proxy: clothing for caste” (Reason 2020). ↩

3. Yarkoni specifically refers to psychology as engaging in “cargocult science” in (2019). ↩

4. One reason intuition and subjective judgment have been given over, so easily, to quantitative measures is the case made by behavioral econ against such judgments. (In other words, the defense of pseudoempiricism comes from pseudoempirical work.) And yet our ability to parse highly indexical social signals, or linguistic expressions, within an *ecological context*, and to an extent that consistently defies our best computational models, testifies to our ignored "indexical genius." As Perry puts it in "Ignorance, a skilled practice": "Overconfidence in the global knowledge game, especially [among the] social sciences, threatens the production and appreciation of the genuine kind of indexical knowledge that humans are geniuses at producing and using" (2020). ↩

5. The connection with legibility: a lack of modesty toward complexity, and a lack of faith in the hard-to-scrutinize. In surrogation, the lack of faith lies in judgment; in legibility, in the forces of cultural evolution and the local emergence of sense-making and self-order. ↩

6. Looking to the most sclerotic and dysfunctional arenas of our society —politics, policing, institutionalized art, among others—suggests that one of the main problems of the modern world is a much softer type than the traditional corruption. Our society is not so much meritocratic as it is *optikratic*: to be seen *is to have* power, just as being seen to have power is also to have it, and power is awarded not on the basis of "actual" (in the ideal sense) merit or value, but on the basis of sporting their appearance. This is at once utterly obvious—one's impression of an object is all one can, finally, operate off, and thus there "is no other way"—and at the same time seriously non-trivial: the translation of holistic quality to public appearance is lossier than usually assumed or acted-upon. ↩

7. Actions like Yarkoni's which alter the common knowledge of the field and thus potentially alter its internal incentive structure, may improve the situation negligibly. ↩

8. There is also the fact that those associations which endure—which are robust to destruction by free-riding—must be built on a logic of *cost disparity*: the public marker must be disproportionately cheap to exhibit for those who really possess the private quality, and accordingly, be disproportionately expensive to those who do not. A middle-class individual might be able to *afford* a Rolex, or a very expensive car, but it so impacts his finances and freedoms, requires such significant sacrifices that it is only rarely and in specific circumstances worth it, in the final cost-benefit analysis, to "purchase" a signal. For the upper-class individual, such expenditures involve very little sacrifice at all, and the small gain in status from parading such luxury goods will outweigh their cost. There are many associative signals that are *not* robust, however; though they may be ridden into oblivion over months or years, they still exert a presence in daily life. ↩