To everyone whose conversations and feedback developed this book: the TIS boys—particularly Hazard, Crispy, Snav, Goes, de Falco, Holtzman, Beiser—but also P. D. Farley & Kevin Focke. And the landlord's daughter, on whose cranberry bog a draft was written.

At different times, useful advice and pushback was offered by: JoXn, Quin Lewandowski, Will Jarvis, Toko Testudo, Dallas Card, Neil Fitzgerald, Spencer Cavanaugh, Ian "Aloysius," Romeo Stevens, David Strider Wallace, and Nico Kaack.

Abridged Contents

For readers who desire an expedited reading schedule, or who find the introductory discussion of surrogation overly basic, the following route is recommended.

Contents

Introduction

> It takes years of training to get a pilot's license, but it only takes a couple of minutes to steal a pilot's jacket and hat.
>
> Norm Macdonald

What does it mean to be a commercial airline pilot? We hope it means something like "is able to safely fly a plane from one airport to another." But for most people in most situations, commercial airline pilot actually means: that person in the front of the plane wearing the pilot outfit. We're all pretty sure that there are several levels of controls ensuring that the symbol "wearing the pilot outfit in the front of the plane" means "able to fly the plane," but it's a very slim minority of people who actually know what those controls are. Seeing "wearing a pilot's outfit" and interpreting it as "safe to fly" is a common, mundane part of life that works without issue.

Many of the people getting on those planes have college diplomas on their walls. Those are symbols, too. What do *those* mean?

"You attended this college." Probably, if the paper is nice and has a fancy seal, but doubtful if it looks like it came out of a printer. "You can navigate inefficient bureaucracy." Maybe a bit better than most, but you might have been aided by overbearing parents or a dedicated admin worker covering for you. "You're an intelligent person who knows a lot about the thing written on the diploma." We've all met enough clueless degree-holders that we're not falling for this one, right?

We can see that symbols stand in for all sorts of real meanings, sometimes many meanings per symbol. We all have different subjective ideas of how well each symbol binds, and we could summon evidence to defend our positions, but we don't typically think of these opinions about bindings as being the same sort of thing. They're just decisions made so frequently that we don't see them as decisions.

What makes a symbol more or less meaningful *in general*? How would you measure that? Any decision that involves creating a framework of some sort, a "surrogation metric" or "meaning rule," is doomed to fail. *Surrogation* is the study of how well those abstract objects bind to concrete meanings: trying to make an abstract object to describe makes about as much sense as building your house out of hammers.

No, what we need are examples. Examples of symbols that very strongly mean a particular thing; symbols that formerly had a particular meaning but got decoupled over time; symbols that have an adversarial, cyclical dance with meaning; symbols that never meant anything at all. Examples upon examples next to each other so we can break out of the habit of seeing each "what does it really mean" question individually, and instead look for patterns.

Surrogation is that book of examples. To say anything more would be to miss the point. This is the study of when summaries mean the same thing as the stories and when they don't. This forward is a summary that has some degree of connection to what's actually happening. Read these stories of what's actually happening and decide for yourself.

Collin Lysford, 2022

Author Preface

> only the fight to recover what has been lost / And found and lost again and again / under conditions increasingly propitious[1]

None of the ideas presented in this book are original to it. In an age of information glut and endless archive, novelty's stock deserves to plummet; synthetic and indexing strategies to reign.[2] This text is a roadmap to what is already known disparately and obscurely across the silos of discourse. It tries to situate their framings, to find tensions and agreements between concepts and claims. It is probably guilty of playing too fast and loose, of eliding important differences and projecting similarities.

The term *surrogation* is chosen to provide a handle for an umbrella of a pattern. To provide two nouns and a verb (*to surrogate*) that help us talk about a quiet force. Several divisions and variations will be named, but these are meant to be taken loosely and provisionally. They are a way to organize a tour, a pretense for exploring dynamics.

The surrogation problem is a kind of alignment problem. In much of contemporary discourse, alignment is sometimes—myopically—thought of as purely a problem of artificial intelligence.[3] It is forgotten that the researcher Stuart Russell, in applying the concept to AI, borrowed the word

1 Eliot, "East Coker," adapted.

2 Early chapters, in laying down a conceptual foundation, cover the least novel ground. Readers hoping for new ideas are advised to consult §6.

3 *cf.* LessWrong-offshoot AlignmentForum, Brian Christian's *The Alignment Problem*.

from economics. But even economics is too shallow a scope. Alignment is an evolutionary and ecological phenomenon, perhaps the defining quality of relations between agents, between parts of a system, not dissimilar from the notion of "fit."[4] Whether we study multicellularity, management strategy, political organization, or lichen symbiosis we are studying an alignment problems.

To maintain alignment, sophisticated agents must monitor, interpret, and evaluate their associates. This practice is pejoratively termed surveillance; this text will refer to it as *reading*. The success of any strategy depends on the actions of others; tit-for-tat is premised on the recognition of tats. A legal system on identifying crimes.

Surrogation is a patterning in how we, as adaptive, learning agents monitor, assess, and interpret one another—of our reliance on symbols, metrics, and metonyms asked to stand for more than themselves. It is the (necessary, inevitable) replacement and conflation of reality with lossy indicators, or of indicators with indicators many-times stacked. It is both the distance, and our collective amnesia to the distance, between some "thing itself" which causally matters, and the various stand-ins we construct or rely on to track it. Each successive layer of removal and synoptic abstraction is an opportunity which another agent may adversarially exploit, by expressing the symbol while lacking the substance. And even aligned agents will find themselves pressed to perform, especially for perverse or obsolete reading schemas—to "check the boxes," save the spirit, and represent truth by presenting literal falsehoods. This process of performance

4 In the Christopher Alexander sense outlined in *Notes on the Synthesis of Form* (1964), though also arguably in the evolutionary sense of fitness.

and information emission—whether intentional or side-effect—will be called *writing*.

We live in an adolescent statistical culture,[5] in which metrics have a hypnotic, "Circe-like" enchanting power,[6] transforming men into pigs and Scylla from nymph into sea monster. Much of the existing research in this area has focused on statistics, data collection, and institutional metrics. The pattern, however, lies at a deeper level of inference, information, and interaction. There is no living outside surrogates, or without surrogates, but the extent of surrogation—the extent of our removal and our amnesia regarding that removal—matters and varies. Modernity demands increased information processing from the position of greater distance. It also accelerates change, destroying the environmental regularities on which all surrogative strategies depend.[7] Surrogation problems will continue to grow more expensive as the scale of our coordination grows.

I wrote *Surrogation* while writing and thinking about the pragmatic, functional dimensions of language; it may be profitable to remember, while reading, that words are one of our most common forms of surrogates, that everything which is said here about institutional performance indicators or menswear is also true of words. Their statistical nature, the brute-associative cognitive capacities they build atop, the treadmills and deceptive strategies that result.

5 As Stephen Holtzman would say.

6 Gioia, "The Circean Transformation From Substance to Image" 2002.

7 See §5.6-5.7 for an exploration of environmental drift. See §6.5 for a discussion of surrogation effects in modernity.

12 Ultimately I remain unhappy with the surrogation frame, which has come, in the writing, to feel more like a middle way than a resting place. There are cracks and incoherences in its paradigm—incoherences which I believe were also implicit in the ideas and "laws" that have lent the concept its shape. Bringing it all together, at a level of abstraction higher than typically presented, lets us begin challenging these incongruities. What is "the thing itself"? At what point do we give up the idea that we interpret a signal in context, and cede gestalt cognition? A better paradigm beckons, perhaps one which uses inference, information, and typification as its basic concepts.

> For us, there is only the trying. The rest is not our business.[8]

8 Eliot, "East Coker"

1. Selection Games

1.1. SINGLE-PLAYER GAMES

The great Arabic scholar Abū al-Ḥarīrī is selecting among rocks to build a wall.1 There are certain criteria he uses to make his selection, criteria which involve a rock's apparent shape, size, and kind. The rocks have no perception of al-Ḥarīrī's selection, nor an interest in whether they are selected. Each is unable to change its appearance in any way to alter its chances of selection, even if it were aware and interested. There is only a static, one-way perceptual relationship: al-Ḥarīrī reads the rock, in determining what action he will take upon the world (Fig 1.1).

al-Ḥarīrī now decides to build a wooden fence, and wanders into the woods to choose a source of material. The trees he selects between have some perceptual awareness of whether they have been selected (there is an abundance of evidents that plants register and react to bodily damage). And they have an obvious interest, if we may use that word, in not being selected. But the relationship stays simple because a given tree cannot adaptively alter its appearance in real-time to morph its odds of selection. If al-Ḥarīrī is searching out a sturdy maple, the forest's maples cannot feign the look of oak trees—nor would they know to.

But though an individual tree lacks the intelligence or bodily agency to adaptively alter its appearance, trees in general are

1 A direct explication of the surrogation problem will take some time to reach; impatient readers, or those who dislike worldbuilding, should skip to "1.8. The surrogation family" on page 37.

Fig 1.1

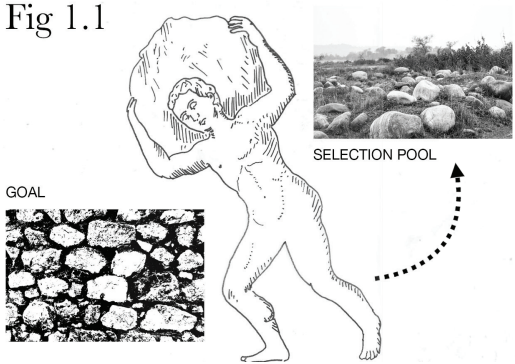
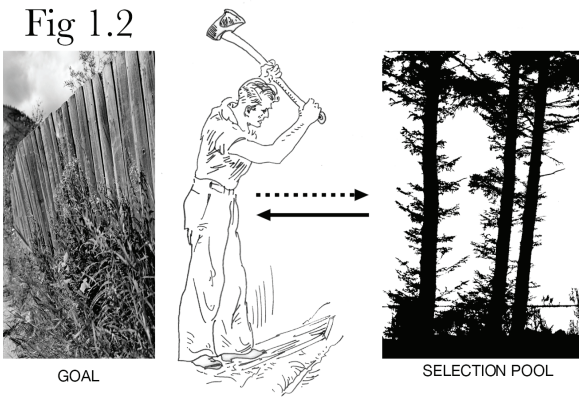


Fig 1.2



governed by evolutionary dynamics which in the long-term amount to a learning process. al-Ḥarīrī inevitably will use certain markers of mapledom—the shape of its leaves, its branching patterns, the thickness of its bark. As a result of these perceptual anchors cum selection criteria, maple trees that have atypical branching patterns will out-survive maple trees with more typical patterns, and over sufficient generations, al-Ḥarīrī's descendents will one day enter a forest of maples which do not look like maples. As a population, the trees have responded to the criteria of the selection process such that none possess those cues which al-Ḥarīrī used to vet his candidates.[2]

At this point, the relationship may be diagrammed as *Fig 1.2*, where the solid gray arrow signifies selection-countering behavior by the trees. While this is a situation in which surviving trees have "learned" to avoid al-Ḥarīrī's selection, many selection games are characterized by an object's desire to *be* selected (e.g. for some preferential treatment, as in mate choice, apprenticeship, or workplace promotion). To say their adaptive moves are "selection-countering" is merely to say that they are adversarial—the result of misaligned interest between the trees and al-Ḥarīrī. al-Ḥarīrī, as the selecting party, is (broadly) interested in discerning the pragmatic truth about his object of selection. He wishes to identify all trees that will (in actuality) advance his fence-building project, and to not waste time and energy on trees which contribute poorly to this goal. His interest lies in a clarity of vision,

2 This dynamic is fundamental to what is known as Vavilovian mimicry. Rye and oats were originally inedible weeds which evolved, through selection pressure, into proper crops. That rye which looked least like wheat was culled—literally weeded out by farmers—so that over, countless generations of agriculture, surviving rye looked more and more like wheat, taking on wheat characteristics until it became a viable cereal in its own right.

an accurate reading. The maple tree's interest, on the other hand, lies in averting selection, in not being felled. We will see this conflict repeatedly, where the selecting party wishes to find the "correct" fit for its search, and the selected object wishes for the selection outcome most in its interests. It is one of the central tensions of selection games.

1.2. TWO-PLAYER GAMES

The dynamics so far explored come to a head when we move to "full-bodied" inter-agent selection games, defined as those games where the object of selection simultaneously (1) has a stake in being selected, and (2) is able to alter, in real time, its odds of selection.

al-Ḥarīrī is writing his *Maqāmāt* in the tall grass when a lion spots him, and the pair enter a selection game for the lion's lunch (with other, spatially and temporally separated prey animals competing with al-Ḥarīrī to avoid selection).[3] It will use certain perceptual markers, such as al-Ḥarīrī's size, physical distance, and gait as proxies to the expected caloric return of al-Ḥarīrī as a meal, whether al-Ḥarīrī is aging or injured and thus easily caught.[4] If al-Ḥarīrī is malnourished, it will behoove him to nonetheless expend significant energy to foster the opposite impression—putting on a show of vitality, or feigning aggression, rather than limping weakly through the tall grass. al-Ḥarīrī, as a full-bodied agent, is

3 It is often forgotten that many of our species' earliest strategy games were against large cats, perhaps our most formidable pre-historic predator.

4 Many predators have precise algorithms, honed by evolution, for how long they are willing to give chase to prey before the expected energy expenditure outstrips expected caloric returns.

able—unlike the rock or maple tree—to fully play out his side of the selection game: he is perceptually aware of being engaged in a selection game; he is able to—perhaps on account of cultural transmission—roughly model the criteria by which a selection will be made; and he is able to adaptively alter the probability of being selected. And, of course, he carries an active and non-trivial stake in the outcome of the game.

A common but naive diagramming of this situation would cast al-Ḥarīrī as writing to the lion, while the lion reads the scholar. This neglects key dynamics: first, that any successful writing is premised on efficacious reading; second and more subtly, that successful reading is analogously premised on skillful writing. A predator carries an awareness, implicit or conscious, that his prey wishes to avoid his selection; rather than approach its prey directly it hides, or sneaks, or sprints. Just as an agent who is the object of a selection game is incentivized to strategically alter his appearance to the selector, the selector is incentivized to strategically alter his appearance to the object, for the purpose of preventing counter-moves. Each party simultaneously assesses while attempting to influence the assessments of the other.

In more adversarial games, writing strives to distort or obfuscate truth; in more cooperative games, it seeks to underline it. A man who has been called up in the draft, whose health is being evaluated for military conscription, may play his hand quite differently depending on his political support for the conflict, or whether he finds it desirous to serve. "Health" is a holistic, hard-to-evaluate, and loosely specified quality, disambiguated only slightly by the specific concerns and stressors of combat duty which gave rise to the spirit of assessment. It must therefore be instrumentalized, for instance

through a check-list of indicators which a credential-holding physician draws up and tallies into a recommendation. The ideal indicators are cheaply, objectively evaluable—they sit on the surface and are easily measured against benchmark. They are also difficult to falsify (in signaling terminology, they are "costly").[5]

Heart rate is one such ideal indicator—or surrogate[6]—of cardiovascular health. And it was therefore not unheard of, during the Vietnam conflict, for combat-wary draftees to dose amphetamines in advance of their physical exams, so as to be disqualified from active duty. Becoming savvy to this writing strategy, the military began detaining individuals overnight who showed abnormal heartrates—allowing the drugs to wear off.

A combat-hungry individual, on the other hand, might strive to conceal or downplay health problems, to avoid disqualification. (This too is undesirable from the perspective of the selector, as such conscripts can become liabilities on the field.) But a "true patriot"—one who takes on the interests of his nation as if they were his own; in other words, who aligns himself with the national organism—will aspire to complete transparency[7] and communicativeness, wanting only the truth as it satisfies the agenda of his selector.

5 Knowledge of the surrogate indicators used, and the way such indicators are interpreted by the selector, is critical in selection objects' ability to shape outcomes. And once known, the surrogate systems develop a gravity of their own—targets to be gamed, in the adage commonly called Goodhart's Law.

6 For now, suffice it to consider a "surrogate" a superset of indicators, markers, metrics, proxies, cues and signals—those telling signs which gesture at a more important, hidden whole.

7 "Transparency" is a problematic metaphor; as we will see, the facts are often anything but self-evident, and must often be performed—or

A desire—particularly acute in large bureaucratic organizations—that decision-making be ritualized, auditable, and routine—rather than dynamic, reflexive, and contextually adaptive—often leads to a simplistic, static attitude toward assessment, and to a simplistic, static evaluative process.[8] Those surrogate-metrics[9] employed tend to be generically, naively, and straight-forwardly implemented, with a premium on public transparency. This makes them "gameable." The gaming of metrics is sometimes referred to as "Goodhart's Law"—that any measure which becomes a target ceases to be a good measure. And yet, it is precisely our lack of understanding of selection games which leads us to see this dynamic as a conditional law—*if* a measure becomes a target—rather than as an inevitable outcome of all evaluation and surveillance systems. (And by extension, of all selection and strategy games.) *Any* measurement or indicator that exerts selection pressure will be targeted by its objects of selection—either in real-time, as in the adaptive intelligences of two-person games, or in evolutionary time, as in the felling of trees. Naturally, this targeting damages the surrogate's efficacy as evaluative tool. When this inevitable and natural player behavior goes against the game's spiritual basis—that is, when it subverts the institutional goals of the game host

"dramatically realized," to use Goffman's term—so that others may recognize them.

8 Routinization and scientism are also decisional anxiolytics for selectors, shifting or deferring the selectors' own judgments (and the responsibility which accompanies judgment) onto some "objective" system of evaluation which both absorbs blame and can be projected upon with a fantasy of unimpeachability. See "4.9. Optiksmization as Cargocult" on page 125.

9 I use "metric" to mean any measurement that is used as the basis of selection decisions (and in being used, exerts selection pressure on, and thereby alters the behavior of, candidate-actors).

or proprietor, while technically obeying its letter of law—we will call it "degenerate play." [10]

1.3. STRATEGY AND SELECTION

We have so far focused on selection games, a specific architecture of strategy game in which candidate "objects" are compared by an evaluating "subject," and are either selected or not selected in a binary way. Often, selection entails a conscription or expulsion of the object into some larger structure which the subject selects on behalf of—as in military conscription, political election, college acceptance, criminal

10 The term originates in early *Magic: The Gathering* communities, referring to both tactics and the players who employ them. It is meant more technically than pejoratively: such play literally causes games to fall apart. Still, the *Dungeons & Dragons* alignment concept "Lawful Evil" bears structural similarities to degeneracy, insofar as evil may be defined as a style of play that fatally destabilizes coordination past the point of repair, and thereby terminates the (aspiring-to-be-infinite) game for all players. In other words, degeneracy and evil are not, in the ultimate reckoning, merely destructive but also self-destructive.

Tabletop gaming communities have developed the folk concept of the "rules lawyer" to describe players who are pedantic or nit-picky about the letter of the law, in a way which degenerates play or violates spirit. To compensate for such player tendencies, many roleplaying games have adopted a "Rule Zero": The dungeon-master is always right. A popular /r/dndmemes comment describes a variation on Rule Zero which beautifully exemplifies the bargaining quality of all voluntary play:

> Rule 0 of D&D: The DM always has the last word.
>
> Rule -1 of D&D: A player can always leave the game, therefore the DM should be prudent in the exercise of Rule 0.
>
> Rule -2 of D&D: It's a lot harder for a player to find a new table than for a DM to find new players, therefore players should be prudent in the exercise of Rule -1.

trials, contract awardings. Prey is selected out of an ecosystem; its cells are incorporated into the body of the predator.

And while our examples have so far been either ecological or ancient, the selection game in its technologically mediated form, has multiplied to become one of the dominant structures of modernity, upholding the liberal order.[11] The cultural equivalent of artificial selection,[12] such games are characterized by an ecologic of testimony over physics, held together by the dynamic solution-finding capacity we call intelligence.

Relative strangers must vet each other across brief windows of mutual exposure. Trial and error in the field is an expensive sorting strategy, best not done blind. Nightclub door policies, mobile dating apps, hiring rounds, and rental applications are prototypal modern selection games, gating alliances, intimacy, and interdependence. They sit in contrast with more casual and informal interaction styles, which while still strategic tend toward more continuous and open-ended outcome space, and often emerge between well-acquainted agents.

Virtually any interaction between agents can be meaningfully construed as a strategy game in the broad superset sense in which "selection games" participates. By definition, agents have goals (desires, preferences) whose pursuit will varyingly conflict and align with other agents' pursuits, and whose attainment is a product, in part, of those agents' actions. These strategy games may be as cooperative as trying to avoid a

11 See "6.5. Close and Distant Evaluation" on page 190.

12 The breeding of plants and animals.

highway collision, or as adversarial as total war.[13] They may be as simple as rock-paper-scissors, or as complex as cryptography. Each player's desired outcome, and his own best moves toward securing that outcome, depends on the actions of other players. It is in his interest to read these players for surrogates which testify to future actions, and to sabotage or support their courses of action through the strategic emission of signs.

1.4. MULTIPLAYER GAMES

In human society, selection games quickly become strategically and relationally complex. History is encoded in the cognitive schemas of players, as well as the letter laws of institutions,[14] just as it is encoded in the genetic instructions of evolved organisms.[15] Functionally, such games are almost always multiplayer, instead of simply two-player.[16] Whereas, in the examples of the maple trees or lion, there is a clear, intrinsic[17] payoff as the result of the selection game—the lion gains or loses a meal, al-Ḥarīrī his life—human social life is marked by *ex*trinsic judgments (as in debate or figure skating)

13 Which, as Schelling reminds us, is never truly total. By shorthand, we will speak of cooperative and adversarial games (as well as strategies), but real games are never "pure," and real combatants always share interests in common.

14 This idea was first brought to my attention by Gianni de Falco.

15 See e.g. "good regulator theorem."

16 Freudian concepts of super ego and introjection, Lacanian concepts such as the Big Other, and Foucauldian concepts such as the panopticon, fill out some of this picture. Undo Undue's short fiction "The Sexual Act" (2022) plays with this idea for comedic effect.

17 i.e. automatically allocated on the basis of physical law.

where observing third parties are tasked with making subjective assessments as to the game winner, and the allocation of payoffs is a result of obedience to social custom instead of physical fact.[18] And in these scenarios, naive models of perception and judgment, which fail to acknowledge the recursive, adversarial nature of selection games, fall short of adequately modeling their relevant dynamics.

As a result, players are not only engaged in games of reading and writing with one another, but also with these third-party judges and referees. Public life, taking place as it does in front of individuals whose opinion is, on average, of consequence to players, turns all two-player games into multiplayer.

Previously we have simplified the goal of the evaluator (be they referee, hiring board, admissions panel, military doctor, blind date) as access to "truth"—for instance identifying only those civilians who are mentally and physically sound enough for military duty. When selection and strategy games become entwined or nested, this simplification misleads us.

18 Goffman (*Strategic Interaction*, 1969) defines an intrinsic payoff such that "the course of action taken and the administration of losses and gains in consequence of play are part of the same seamless situation, much as in duels of honor, where the success of the swordsman's lunge and the administration of an injury are part of a single whole." It is specifically the extrinsic nature of incentive structures (or internal games, or socially mediated reward structures in general) which makes them "optikratic" (that is, based on outward appearances as much or more than merit). "A clear hit in mortal swordplay can perfectly well occur in a foggy night, the clarity of the hit having to do with its psychological consequence for the hit organism. But in games like fencing where hits are merely points, a move must often be terminated with an act of perceptual clarity, lest there be a dispute as to what, actually, happened."

Establishing shot: the interior of a police station. Detectives have brought in two suspects for questioning. One is suspected of murder, the other of being an accomplice or at least a witness to the killing.

The basic structure of the selection game is this: The suspects wish to escape a legal conviction, and preferably, a court appearance. There are a set of formal rules (see "3. Formal Games" on page 59) concerning what constitutes admissable evidence, and what kinds of sentences will be applied if a defendant is found guilty; although the judge and the jury[19] introduce human discretion, their judgment is still guided such rules. (The jury is asked not whether the defendant ought to go to prison, but whether there is overwhelming evidence that the defendant committed the crime, in which case—when the selection object is found to "match" the selection criteria—he is selected for sentencing.[20]) Meanwhile

19 The jury, of course, is picked through a selection tournament, whose dynamics are described in detail by Christina Marinakisin in conversation with Zachary Elwood (2018). Although jury duty is mandatory in the United States, exemptions are built in to minimize both juror hardship (a form of mercy toward candidates) as well as bias (a method improving jury outcomes). Citizens tend not to want to be picked, and will often exploit whatever exemption criteria are available to escape jury duty—hence, the very mercy of the selector makes it more exploitable. This is not to sing praises of the American legal system but to point out that in many selection games, selection is undergone with an ethical bent—an intent of minimizing harm, and of fair application—which also makes the game more exploitable. (See also disability accomodations in standardized testing). This is also, roughly, the logic by which sentimentality is discouraged in spy films. Mercy makes the merciful vulnerable.

20 In other words, *ought* flows readily from *is*; see "3.4. Decision Rules & Magic Words" on page 72.

the detectives (roughly speaking) can be modeled as desiring to identify and earn a confession from the actual perpetrator of the crime, while operating within formal rules as to how they can obtain evidence or a confession.

The detectives first tell their primary suspect that his friend is cooperating with the investigation: that they've provided him with lunch he's been so helpful. They then parade the friend past the interrogation room with a Happy Meal in-hand—the friend having no idea he is being used in a ploy, confused why the detectives have been so friendly. But the detectives are attempting to strategically misrepresent the game state in order to provoke a confession, and use the McDonald's meal as a "confirming" metonym to reinforce their (mis)representation. Taking cues from Sarah Perry's "Puzzle Theory"[21] and Emanuel Schegloff's work on conversational interpretation,[22] we can say that the initial appearances of a surrogate (when taken provisionally, that is, non-naively) *alludes*, *suggests*, and *implies*. Once an interpretation has been suggested, it can be "confirmed"[23] by later signs which would be predicted out of (i.e. are more likely given that) the suggested theory (is true).[24]

Whether or not he is responsible for the murder, the suspect's interest is in self-representing himself as innocent, or not worth pursuing legally—that no evidence can stick to

21 *Ribbonfarm* 2015.

22 *American Journal of Sociology* 1996.

23 Of course, as in the similarly structured scientific investigation, confirmation is never final.

24 This dynamic underlies linguistic interpretation, where previous utterances are regularly being confirmed, contradicted, or retroactively re-interpreted given later context.

him. The detectives, meanwhile, are trying to manipulate his assessment of the situation so that he commits a game-forfeiting blunder. The situation is, in its fundamentals, not so different from al-Ḥarīrī up a tree after being chased by a lion, watching the lion wander off. He now believes the coast is clear and comes down; meanwhile, the lion has snuck around back, and pounces. Actions are based in perceptions, and by manipulating perceptions, one can manipulate opponent behavior to the opponent's disadvantage.

Next, the detectives bring the suspect into a room with a Xerox machine, and tell him that it is a lie detector. Again, they are manipulating (his impression of) the game state, here by presenting a false front. A sergeant pretends to be a "professor" in charge of administering the lie detector, which is "never wrong." The sergeant's false identity works in part because he's wearing metonymic suspenders.[25] The detectives then pull a scam on the suspect: they strap his hands to the glass of the copy-machine, have it first print "true" when asked his name and address, then print out the word "false" when the suspect is asked, and responds in the negative to, whether he committed the murder. At this point, convinced he is beaten, the suspect breaks down and confesses. (The corollary of Ennius's "The victor is not victorious

25 This scam would be difficult to pull off against an affluent, college-educated person. It is precisely the low fidelity of the (poor, under-educated) suspect's stereotype of academia that lets such a crude, Halloween-costume imitation of professorship work. Stereotypes we can understand as a rough character profile which implies a general operating procedure, making players more legible and thereby predictable to one another. We use composites of surrogate markers to identify types, e.g. *glasses, bangs, sundress with bird print* for *indie, Zooey Deschanel-type*; see "4.4. Typification" on page 103. Stereotype fidelity is higher for in-group adjacent cultural roles. For an extended discussion see Hotel Concierge, "The Tower."

if the vanquished does not consider himself so.") The real selection game they are playing here has been obfuscated for another sort of selection game. The suspect believes his confession is irrelevant to the game outcome because the detectives have already proved his guilt through the Xerox machine, as well as securing a confession from his accomplice. The countering strategy which would prevent his being selected for imprisonment—the silence which, sans material evidence, would preclude his conviction—is made to appear unavailable or fruitless as a strategy, and so he fails to use it to counter the detectives' attempt at fishing out "the truth."

We can get now to the problem of saying that the detectives simply wish to identify (select) the actual murderer. There may be some intrinsic reward, for a detective, to catching "the right guy"—a satisfaction which exists regardless of whether anyone else knows, regardless of whether he receives a financial bonus for his work or is applauded by colleagues. But note that this "intrinsic" reward is still built solely atop the detective's conviction that he has caught his killer—that *his* man is in fact *the* man. Belief, not reality—and we are all now familiar with the extent to which our perceptions and theories are desire-motivated, the extent to which we are capable of fudging the data to convince ourselves of some convenient "fact."

Meanwhile, the larger incentive structure which our detectives are embedded inside—the symbolic and literal capital that is distributed conditional on their performance—is a tree of socially mediated, appearance-predicated selection games, top to bottom. The extent to which such a system optimizes for truth is the extent to which it is rigorously constructed towards cross-examination, oversight, skepticism, checks and balances—and the extent to which the system's

constituent members are dedicated to identifying truth, in their own actions and in others. Such dedication will not emerge on its own; it must be rigorously screened and selected for in entrance games.

Such screening and selection procedures come to define institutional composition. Economics has analyzed many failure modes of organization and collective action: conformity, risk-aversion, asymmetrical justice, preference falsification. But one aspect, somewhat less discussed,[26] is far more crucial. Insofar as an institution is a body of individuals, with varying capacities as decision-makers, varying ideals of integrity, communicative capacity, and coordinative inclination, it is the selection game—the assessment which qualifies an outsider to serve within an institution, or an insider to climb the ranks of power—that counts most.[27] Selection games are the screening mechanism which keeps eccentric talent out, or mistakes glittering image for actuality; constructs a cycle of accreditation, or a pseudoscience out of psychology. Rules

26 Although see e.g. Stiglitz on matching games, and the field's adoption of Lewis's "signaling game" concept, for related work.

27 For instance, there is still power in that near-tautology (quasi-evolutionary) that in order to win a selection tournament, one must become the sort of object that *can* win the selection tournament—that there is a "shape" which one must be or become to pass through the tournament, like a lock and a key. And so when a player (perhaps a politician, corporate executive, or MFA student) who has survived one of these tournaments, and succeeded by what are basically conventional tactics, appears or claims to be a novel presence in the organization, it is unlikely that this difference is more than skin-deep. Only if the tournament has been won in an unconventional way do we have the possibility of real novelty in the composition of the system. On the admittance of Jewish and Asian members to a (traditionally WASP) California Bay Area country club, Nick Greer (2023) writes: "These outliers are never actually outlying, but have assimilated into the fringes of this culture, often through a precise performance of the in-culture's aesthetics and customs."

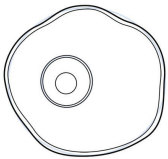
and culture, the "structure" which is more regularly blamed for the shortcomings of bureaucracy, are determined by early membership selections, mere byproducts of the org's first selection games.

Superorganisms, from body cells to human institutions, tend to have both a hierarchy of decision-making power and a nested structure of boundaries, which prevent the entrance of toxins or bad actors (in short, things which work against the superorganism's goals) while admitting goal-furthering resources and subagents. These nesting boundaries are maintained by selection games, upper levels of hierarchies continually monitoring the behavior of lower levels for selection out or promotion upward.[28]

The relationship between the detective and the suspect is, within the present framework, fundamentally like the relationship between the detective's supervising officer and the detective: each are involved in a layer of selection game in which they hope to "look good" with respect to the selection criteria (be it towards competence, in the case of the detective, or innocence, in the case of the suspect).[29] The

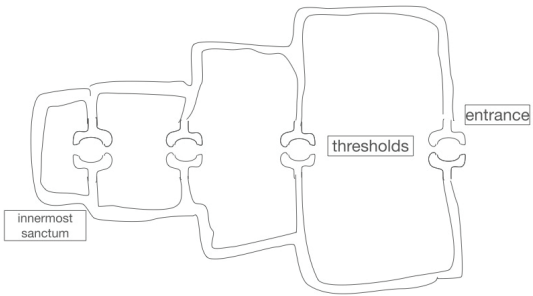
28 Insofar as an institution is a body of individuals, possessing varying capacities as decision-makers, varying ideals of integrity, communicative capacity, and coordinative inclination, it is the selection game—the assessment which qualifies an outsider to serve within an institution, or an insider to climb the ranks of power—that counts most in an institution's overall performance and quality. Selection games are the screening mechanism which keep eccentric talent out, and mistake glittering image for actuality; which construct a cycle of accreditation, or a pseudoscience of psychology. Rules and culture, the "structure" which is more often blamed for the shortcomings of bureaucracy, are determined by early members, even as they are merely an influential byproduct of the organization's first selection games.

29 And indeed, the "professor" of the Xerox scam is the unit's sergeant; by including him in the routine, the detectives get to show their



left
a simplified drawing of an animal cell

bottom
Christopher Alexander's diagram of temple architecture



same is true from the detective's supervisor up to the city police commissioner and beyond. Even at top levels of both private and public sectors, senior officials are still answerable to Congress, courts, shareholders, the public, etc.[30] It is selection games all the way up and down. The detective desires promotions, capital, the respect of his colleagues, and to avoid being fired or prosecuted for his actions. And because the "truth" of his quality as a detective is radically under-instrumentalized, and can never be known with anything ap-

(selection-empowered) superior that they are at work, that they are effective at their jobs, etc.

30 This, perhaps, ought to shift our intuitions from attributing blame for corporate behavior or ethical breaches *away from corporations themselves* and onto consumers, who act as selectors on which corporations survive, thrive, or perish.

proaching certainty but only testified to via publicly available signs, his is a game of appearances.[31] The detectives who appear competent will (to simplify) be promoted—just as the suspects who appear innocent will be let off. Extrinsic, multiplayer games are doubly ruled by appearances, because appearances are all that are available to the agents passing judgment on their outcomes.

1.6. INTERNAL GAMES

Values—an agent's deep, "theological" priorities and goals—are difficult to change. A man is not easily persuaded to give up his family, his religion, or his country. But values only partially guide player actions; our assessments of what *is* are as crucial as our beliefs about what ought to be. (As the saying goes, if you want to get a good person to do heinous things, make him believe the cause he is pursuing is just. Not to change his sense of justice, which is difficult, but to change his understanding of action consequences so that they align with his pre-existing concept.[32]) Employers are less likely to be persuaded to value incompetence in their employees than they are to be persuaded that a given interview candidate is competent and therefore deserves hiring. A prospective

31 Politicians have such perverse, optikratic incentives because voters are so distant from them—they lack real, ecological proximity, and receive only a narrow bandwidth of doctored, public relations work on which to base their selection decisions. This of course selects for candidates primarily on the basis of their public relations effort, and not their actual competence or ethic. See "4.8. Optikratics" on page 118.

32 When individuals discover or believe this has been done to them—that a government, press, activist organization, advertiser, etc has framed a situation strategically in order to provoke a set of corresponding actions—they say they have been "played."

employee is less likely to be persuaded that his priorities include the success of the company, in its ongoing outer game of maximizing shareholder value, and more likely to enlist his help if, in doing so, he receives returns according to his own pre-existing priorities. That by playing on the company's behalf, he might provide for his family, earn status, and develop marketable skills for future employment.

Organizations must therefore erect (or else discover) an incentive structure which artificially or automatically doles out player-desired payoffs in exchange for organizationally desired behavior. We will call this the inner (or "internal"[33]) game of interactions. And while one can strategically self-represent in a deceptive way so as to secure such extrinsic payoffs, in a well-designed incentive structure it should be less expensive on average (in time, effort, cognitive load) to simply enact the payoff's prerequisite in actuality.

Players typically enter constructed incentive structures voluntarily—one wishes for the rewards of a job, and therefore enters its internal game in hope of securing them. Involvement in an internal game is typically initiated by an entrance game characterized by mutual matching, as in fraternity rushing, job applications, or college admissions. The institution which, should the entrance game result in a match—mutual selection—will host future internal games, acts similarly as host of the entrance game, and the entire game is frequently marked by an implicit "narrow and choose"[34] algorithm:

33 To avoid confusion with the concept of "inner game" popularized in such publications as *The Inner Game of Tennis*.

34 What I call narrow-and-choose games are simply those in which two or more parties coordinate to make a decision by alternately whittling down the set of possible options through elimination rounds. Simple versions are only two rounds, e.g. a player might begin by suggesting a short list of restaurants

applicants select which entrance games are worth applying to; hosts decide which applicants are worth admitting, extending offers of acceptance; and admitted applicants are allowed the final choice among admitting hosts.

There are notable exceptions: Individuals do not voluntarily enter the legal structure of their birth society, and must instead opt-out (an often costly decision).[35] And individuals, if they have lost certain selection games in the larger legal structure, risk forfeiting their right to continued consensual play (military drafts, mandatory education, imprisonment).

Because it is difficult to alter the abstract priorities or "values hierarchy" of optimizing organisms, regulatory structures built to facilitate inner games often yield what in artificial intelligence research is called the "nearest unblocked strategy" problem. Patches to the inner game's incentive structure do not alter underlying player motivations but merely erect one more roadblock around which the agent routes in pursuit of its previous goal. This is famously the problem of centralized economies, and the advantage traditionally attributed to market economies: capitalism as a system which brings into alignment otherwise mis-aligned self-interests.[36]

he is willing to dine at, or films he is willing to watch, and allow a companion to select any item from that set.

35 The same is true of the (albeit informal) family structure.

36 Capitalism in its Hayekian formulation can also be considered an extension of standpoint epistemology; see Matthew McKeever's treatment, "Capitalism Is A Standpoint Epistemology" (2018).

Alignment itself—too often reduced to merely an economic or AI safety problem—is perhaps the fundamental problem of complex life, as well as one of the most difficult and profitable (in a sense that far supersedes financial profit) problems to solve. Ultimately, it can only be solved locally, drawing on the concrete affordances of a situation. But a few general principles might help guide the search for local solutions.

What I think is becoming clear, with our notion of selection games, is that representation undergirds alignment, particularly (but not solely) in non-evolved systems. Natural selection, in the longue durée, tests the real, while intelligence infers from the apparent. In between lies a world of difference.[37]

This reliance on representation is true even in the ecological huddle of hunter-gatherer life. There is no doubt that stories, symbols, and signs played a strong role in individuals' reputations, just as they play a prominent role in the animal kingdom. But the reliance on representation is pushed to an extreme in modern life, on account of its increased complexity, its technological mediation, its abandonment of localism for globalism, and its urban populations of increasing anonymity—all of which demand, in turn, increasingly synoptic (and therefore, increasingly removed) views of the actual.

Alignment is expensive to monitor and oversee, and has an upper-bound of visibility; there are certain realities hidden to outside observers, or even internally, to the conscious mind

37 See also the PvE and PvP distinction.

itself.[38] In fraud detection, there is a motto that zero fraud is not the optimal amount of fraud. It nods at the economics of preventing fraud—the cost to oversee an economic system so thoroughly would far outstrip the loss from occasional fraudulence. There is a frontier of diminishing returns, or increasingly expensive tradeoffs, whereby catching increasingly marginal amounts of fraud becomes proportionally more and more expensive.[39]

The situation is similar in alignment; we may think of deception (or synonymously, in an information context, defection) as fraud. Strategic representation can create the appearance of alignment where none exists. Fertile ground for alignment can be overlooked because interests are poorly represented or misunderstood. But it is far cheaper to rely on lossy representations for oversight, and so we accept the occasional false negative or positive as the price of convenience, the price of making inferences on the inaccessible. One implicit contention of the present text is that we are often naive about this price, particularly how quickly this price inflates in a rapidly changing environmental context. That we underestimate the extent of false positives and negatives, and the cost of their entrance into, and promotion up the ranks of, institutions (and into positions of selection-guiding power). That this underestimation is a primary—but often overlooked—source of the corruption and inefficiency attributed to modern bureaucracy. That the wrong surrogate, in the right place,

38 *cf* the work of Robert Trivers on self-deception.

39 This baked-in notion of inevitable tradeoff—where in the pursuit of maximizing a given property or outcome of a system, each marginal gain comes at increasingly high costs to other properties or outcomes—is, I believe, inherent to the notion of optimization, although I am as-of-yet unaware of a formal proof.

can tear a society apart, can cause centuries-old systems to crumble.

Much has been written on why "agile" start-ups reliably outcompete more ossified bureaucracies, but the greatest factor might be that start-ups are frequently populated by true believers,[40] whose ideological commitments in combination with high equity stakes make them tightly aligned with one another and with the interest of the start-up. Moreover, these early employees have been selected directly by company founders, who have often had long personal or professional relationships with the employees previous to hire. "Bloated" institutions, on the other hand, lack the intrinsic alignment structures of equity and the spirit-preserving bonds of friendship. They are often populated by hires of hires of hires, and must institute rote, ritualized evaluation methods to monitor employee quality. And while the institutional cost of employees feigning qualifications, or feigning to work, is obvious, more subtle and more damaging is the all-too-common phenomenon of employees working on the wrong thing—expending effort and energy on internal games that fail to advance the goals of their wrapping "organism."

We'll gain a second understanding of the internal vs. external game, and selection dynamics generally, in later sections on mesa-optimizers ("5.6. Mesa optimization" on page 154).

40 Meant in a similar sense to "true patriot," §1.2. So far, this text has often simplified affairs by assuming that the "team" a player plays for is, firstly, himself, then secondly his close family and allies. But this need not be the case—humans are capable of, and regularly do, play on behalf of abstract ideas, or complex superorganisms—and this arguably is the case of true believers.

1.8. THE SURROGATION FAMILY

I have dragged us now through several sections without a straight explanation of the book's titular concept.

The "surrogation" idea came after repeated encounters with ideas and arguments from various fields, all of which, I felt, were connected by a broader set of patterns, a family likeness. Wittgenstein famously writes in his *Philosophical Investigations*:

> There is no characteristic that is common to everything that we call games… It is a family-likeness term. Think of ball-games alone: some, like tennis, have a complicated system of rules; but there is a game which consists just in throwing the ball as high as one can, or the game which children play of throwing a ball and running after it. Some games are competitive, others not.41

At this point, we might say, I had come across the concepts of *football*, *mahjong*, *arcades* and *competitive eating*—but I lacked the concept of *game*. The handle *surrogation* is as an attempt at a "game"-like level of abstraction and category, defined by family resemblance more than succinct, necessary and sufficient conditions.

Briefly, the apparently related concepts which I had stumbled upon in the course of other research, and which I liken to mahjong or football, include: From the field of artificial intelligence, *wireheading*, *underspecification*, and *nearest unblocked*

41 Incidentally, it is my belief that modern ecological and game-theoretic models have clarified quite a bit what constitutes a "game"—albeit moreso in the technical usage of the word than its everyday sense. See also Appendix I, "What's In A Game?"

strategy; in philosophy, from C. Thi Nguyen,[42] the ideas of *gamification* and *value capture*; in statistics, those of *overfitting*, *latent* vs. *manifest variables*, *proxy measures*, *Fisher information*, and *operationalization*; in medicine, the *surrogate marker* and *surrogate endpoint*; in psychology and psychometrics, *construct* and *test validity*; in ethology, *signals*, *cues*, and *mimicry* (e.g. *Batesian*, *Vavilovian*); in sociology, *goal displacement*, *legibility* and *Campbell's law*, and to Bourdieu, *capital*; in microsociology, the *symbol* and *symbolization* process; in economics *Goodhart's Law*, *the Lucas Critique*, *perverse incentives*,[43] *attributes*, *signaling games*, and *screening games*, as well as the distinction between *private* and *public information*; in information theory, *joint entropy* and *mutual information*; in metascience, Tom Griffiths' *idolatry* [44]and Feynman's *cargocult*;[45] in games studies, *degenerate play*; in business and military strategy, the *McNamara fallalcy*, the *indicator* (as in *key performance* and *key risk*), and Venkatesh Rao's *gollumization*; in sports, *stat-padding*, *flopping*, and *empty stats*; and finally, in the folk theories of ordinary language, concepts like *fetish*, *masturbation*, *cobra effects*, *cheap play*, *surface compliance*, *teaching to the test*, *what's measured is managed*, and *winning by technicality*. The second appendix gives a brief overview to some of those concepts, which are not otherwise directly treated.

What connects these ideas? The answer is necessarily long and digressive, will take time to answer—after all, inherent in this structure of family resemblance is the lack of any genetic "essence" which can be compressed into a single patter. Each member of a family is related to others, but they do not

42 2020.

43 See also Ivan Illich's *paradoxical counterproductivity*.

44 2016.

45 1974.

all share the same green eyes and red hair. There may not be a single family trait which all share, and even if there were, it would not define them—many non-relatives, after all, would share that trait as well. (As Plato's featherless biped well illustrates.[46]) Still, we can gesture toward the rough strokes of similarity before elaborating details and complications, examining case studies and comparing circumstances.

They are all, at their core, information and inference concepts; many capture a sense of once-removal or representation, premised on a tacit distinction between some "thing itself," desired to be studied or optimized, and some surrogate which for myriad reasons must stand in the thing's place. The surrogates provide information about their surrogateds by virtue of their statistically correlating (that is, their frequently co-inciding). This statistical correlation, which is at the root of the costly signal concept, is often, in the human realm, upheld by systems of surveillance and management—for instance, by the legal ramifications of impersonating an airline pilot. In cooperative games, both parties actively strive to maintain and leverage these statistical correlations in order to understand and be understood, a pattern of behavior with structural similarities to Thomas Schelling's description of focal points in *Strategy of Conflict*. In adversarial games, however, frequentist approaches to meaning are vulnerable, and better replaced by causal explanations. (See "3.3. Surrogate metrics" on decoupling and drift.)

Many of the situations which these concepts describe or emerge from are game-like, with agents competing for limited resources or preferential treatment—although some of them are closer to a single-player, non-reflexive structure.

46 Diogenes Laërtius, *Lives and Opinions of the Eminent Philosophers*.

Of these concepts, most implicitly rest, if only roughly, on a distinction between the spirit and letter of a game—that is, between the desire or intention of a game's designers, and the actual specification of its payouts.

Many of these concepts nod toward the tendency of players to strategically appear cooperative with their superorganism's mission, or with fellow players, while in reality playing for selfish advancement—free-riding on technicality, reaping the rewards of a system in a way that "the system," if it could be anthropomorphized—or a fellow player, if present—would condemn as exploitative.

In some sense, a system has neither desires nor intent; its full character exists in its present form, as it is programmed.[47] There is nothing "beyond" its specification in letter. There is no "hacking" the laws of nature; everything is "in bounds." "Wireheading," reified as something aperspectival, is a teleological misnomer—there is "no such thing," from an objective third-person perspective on a system. The system has no opinion on whether a strategy of internal play is fair, beyond what its literal rules allow or disallow.

The agents that design, uphold, host, and participate in systems do, however, have intentions and desires, which includes a proper "way to play." To those who administer such internal games, this proper style of play is (prior to surrogation confusions) exactly that behavior which motivates the

47 On the other hand, it is not entirely clear that agents behave any differently under the hood, and are not merely complex systems in the same sense as an institution. "Desire" and "intent" are abstract shorthands, and should not be reified, here, as more than that.

creation and administration of the internal game to begin with.[48]

Wireheading—and to some extent, therefore, surrogation—rests on a perspectival interpretation of gameplay. The perspective can come from the game's designers, its audience, its players, or any other entity invested in the game's outcomes. There is no "correct" or authoritative interpretation—only cultures of play and tacit coordination styles,[49] as we will see

48 The motivations behind such games are holistically bundled and endlessly complex, but to simplify for the sake of example, the purpose of the legal game is to maintain law and order, and uphold justice; the purpose of a corporation's internal game is maximizing profit. Play which escapes punishment, or evinces reward, while violating these founding principles is undesirable from the perspective of those who designed and continually maintain these inner games. Game hosts and designers will therefore attempt to legislate it out through continual updates to the letter of law (and thus the structure of formal payout). This subject is explored more thoroughly in §2.1-2.3 (beginning p. 44) and §5.1-5.4 (beginning p. 127).

49 The same is true for the game of literary interpretation: on what basis can one claim author intent, or audience interpretation, is "the" meaning of a text? There is no such basis. It is a verbal dispute, ended if we divide-and-conquer "meaning" for "intent" and "interpretation." On what possible grounds is one or the other the "true" meaning, when native speakers and subject experts differ wildly in usage?

And, importantly, since the interplay between authors and critics (and literary historians, canonizers, the lay public, etc) can be described meaningfully as a selection game in its own right, we should understand that the culture of acceptable interpretive play—what speculations are considered in and out of bounds, in the internal competition between critics, publishers, and their audiences—will have an effect on how writers write, since it has an effect on how they are understood, evaluated, and, essentially, selected. If an author knows his biographical background will be heavily considered, he may misrepresent it, or leave certain connections implicit; if he knows his intent will be foregrounded, then he may go to great lengths to make that intent explicit in interviews, or even come to rely on such extra-textual comments as a crutch

soon enough in David Sirlin's concept of a "scrub."[50] Literal wireheading rests on a system of ethics or theology; without such a system, there is no right or wrong way to use our neurotransmitters; evolution does not have desires.[51]

Surrogates are at their most powerful in selection games, but are a fundamental component in all strategy games, where they serve as the basic unit of inference for players engaged in reading and writing each other. The basic conditions for the strategy game are necessary preconditions of the social: ecological proximity and outcome interdependence. And once we have established and argued for the ubiquity of the strategy game, the presence and ubiquity of surrogative phenomena no longer needs arguing, but may be elaborated on freely. We do not need a law like Goodhart's to tell us that statistical or management metrics will become behavioral attractors—just as we do not need a law to tell us that the

---

in "encoding" the text. (See also the relationship in the visual arts between conceptual work and "explainer" wall text.)

50 "5.4. Sirlin's Scrub" on page 143.

51 We may feel differently—that there just is something troubling about "plugging into pleasure" and neglecting productive social life (as in Nozick's experience machine thought experiment, which surveyed individuals regularly turn down). But this is a cultural value judgment more than a distinction implicit in our chemical reward system.

We can, however, still call such behavior degenerate in a strict sense: pleasure separated from sexual reproduction degenerates the very game which gave rise to wireheading play, bringing it to a conclusion through the extinction of an ancestral line. It is a self-defeating, self-destructive, "evil" play style (in the strict, game-theoretic sense of "evil"). Certain styles of play, certain coordination equilibria, and the letter laws or behavioral norms which facilitate them, are self-sustaining; others flare out in dramatic fashion. Organisms that survive millions of years of natural selection are precisely those whose reward functions regulate their evolution in a sustainable way.

"soft," informal indicators of a job interview are behavioral attractors for interviewees, or that in the long run Abū al-Ḥarīrī will enter a forest of maples that do not look like maples.

In formal, institutional games, surrogation begins with the necessary translation of spirit into letter, often accompanied by an amnesia that this translation has taken place, so that the letter is reified as the purpose itself of the game, rather than being taken contingently as a flawed if useful means of tracking and motivating player behavior. But surrogates are also employed in more informal selection games: cocktail parties, gallery openings, military battles, and children's games. And in both formal and informal games, selectors and candidates, evaluators and evaluateds, alike present a front—a public-facing system of surrogates designed to accomplish a desired outcome, which only correlate with the reality of the presenter.

2. Spirit & Letter

2.1. MIDAS

When an institution wishes to set up an internal game, it must convert a desired spirit of behavior into a specified letter of law.

Spirit takes many forms—rarely can we establish its exact throughlines; we may recognize when we see it, but only in person, in the fullness of its situation. Phenomenologically, it lives at the level of feeling.[1]

Letter—the specification of spirit—can attempt to capture some of the shapes and guises in which spirit manifests, but it will never succeed in full.[2] And yet spirit cannot be legislated, cannot be uniformly instituted as expectation, cannot tile itself across a superorganism.[3] Socrates' interlocutors, like

1 Such a model of feeling as computational is suggested by e.g. Peli Grietzer's work on vibe and Gary Klein's work on expert intuition, but some version of it is implicit in much of psychological and therapeutic practice. As Justice Stewart infamously proclaimed in the obscenity for Louis Malle's *The Lovers*, "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description [of pornography], and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that." The gestalt-cognitive view these models gesture towards is a challenge to the more simplistic sign-context framework developed in the present text.

2 In this sense, game spirit is much like a family resemblance theory of concepts, and suggests a neural net-like emotional-cognitive structure that is statistical, clustering, and associative.

3 From here on out, by "superorganism," I will mean a cooperative enterprise organized by a prestige economy toward a common purpose. This

their analytic descendents, were unable to formulate an elegant, robust specification of the good, the just, or the pious. What, to an institution, to a justice of the court, to a human resources department are such vague senses of extensional matching as these? Insofar as spirit is tiled and legislated, it defies accountability, relies on trust and discretion, precludes routinization and monitoring. So we inevitably retreat to letter.

Letter invariably fails to capture spirit, and this destabilizes the functioning of laws, requiring successive rounds of revision. Each ontological-linguistic failure by game hosts to adequately model the operational reality introduces—is (temporarily) concretized as—strategic opportunities in the field of play. Each rounding off, each unaddressed patch of possible behavior, each synopsis or compression which evokes highly variable patterns of inference and interpretation, shifts the incentive structure, and thereby the behavior of players, away from the desired spirit and toward some other, emergent, perverse or unintended end. Often in this process, the substitution of letter for spirit is itself forgotten; there is an often rapid, coordinated forgetting that something complex and preverbal has been surrogated into language (into measurement, into lossy representation schema) and the surrogate comes to stand in as spirit itself. Not as means, but purpose; not as proxy, but as "the point." Prohibited moves gain a patina of immorality. Rules which were originally instrumental are de-instrumentalized; policies that were context-sensitive lose context. A tactic or behavior is no longer problematic on the grounds that in certain settings it risks an

---

prestige economy is the "internal game" which motivates individual players to coordinate and advance the interests of the superorganism within the "external game" it is embedded inside.

undesirable outcome—rather, the behavior is itself undesirable "intrinsically."[4]

Dionysus promised to fulfill any wish that Midas desires, and Midas wished that all he touched turned instantly to gold. We know or think we know what Midas was "really" after—voluntary conversion, for one—but did even *he* know, precisely, what he wanted when he wished it? Could he have put it into words, specified the parameters and constraints? Or did he assume that a human-like intelligence—with theory of mind, a generous interpretive spirit, and a cooperative bent—would infer the spirit of his desire? Unfortunately, the total space of intelligence is much broader than the space of human minds: Midas was met by a trickster god, and his sloppy specification of spirit into letter became the undoing for which he is known. In some versions of the story, as soon as the wish is fulfilled, the king's sandals and toga transform to gold, and he is encased alive inside a metal suit.

Even a young child, in proposing a fantasy game of "three wishes from the genii," will proactively specify such cases as "no wishing for more wishes," "no wishing for infinite powers," etc. These cases are informed by experience, far more than reason: they are an inventory not of all possible wishes (moves) that would contradict game spirit, but of historically common moves. And if we are playing the genii game ourselves, and decide to wish for immortality, we may be careful to specify a conditional immortality—to avoid eternities of suffering, of being trapped for long periods in an iron casket in the sea.[5]

4 See also §"4.5. Fetishizing Means" on page 105.

5 See also *The Monkey's Paw* for a literary treatment.

2.2. LETTER'S LIMITATION

Quaker epistle, 1656, echoing 2 Corinthians 3:6:

> Dearly beloved Friends, these things we do not lay upon you as a rule or form to walk by, but that all, with the measure of light which is pure and holy, may be guided; and so in the light walking and abiding, these may be fulfilled in the Spirit, not from the letter, for the letter killeth, but the Spirit giveth life.

Why did the trickster god give Midas the perversely literal version of his wish? Was he unable to understand the spirit of the king's desire, lacking a model of the king's psychology, or else a goal-oriented view of communication (as is often the case with artificial intelligence)? More likely, in the symbolic logic of the myth, Dionysus understood perfectly but chose not to cooperate in a game-theoretic sense.

This is key: adhering to letter is a form of cooperation, but adhering to spirit is doubly so. For while adherance to letter, and particularly the adherence to letter while under surveillance, is often to coordinate only and exactly to the extent required to "stay in the game"—to continue playing, and not be expelled or disqualified—adherance to spirit, on the other hand, may go far beyond this. Game hosts frequently lack the authority (or the desire, in order to maintain legitimacy) to expel players on the basis of spirit violations.[6]

6 There is also, occasionally, a selfless nobility to degenerate (i.e. letter-observing, spirit-violating) play, particularly in team sports. Many avoid degeneracy less on moral grounds and more in order to save social face, as social sanction is one of the primary ways that degeneracy is dealt with and disincentivized. To accept social sanction (for instance, being perceived as a less honorable or esteemed player) in exchange for greater team success is sometimes a lauded act. See "5.1. Spirit, Symbol, Reality" on page 127.

48 Harold Garfinkel's "et cetera clause" (1967) makes this more clear:

> No matter how specific the terms of common understandings may be—a contract may be considered the prototype—they attain the status of an agreement only insofar as the stipulated conditions carry along an unspoken but understood *et cetera* clause [...] Therefore it is both misleading and incorrect to think of an agreement as an actuarial device whereby persons are enabled as of any Here and Now to predict each other's future activities. More accurately, common understandings that have been formulated under the rule of an agreement are used by persons to normalize whatever their actual activities turn out to be.

Explicit coordination is never purely explicit; tacit coordination is required to put explicit agreements into effect, and for participants to understand the desired or acceptable implementations of agreements. Even explicit contractual terms, agonized over by high-paid Harvard lawyers (or their freshfaced assistants), remain irrevocably fuzzy and ambiguous. As the problems of Constitutional interpretation make clear, any meaning may be destabilized, at any time, by a shift in its interpreter's cultural or pragmatic context—and it is incoherent to imagine a language which could work otherwise. Hans Vollmer, commentating on the et cetera clause, writes:

> Garfinkel's comments on the "et cetera clause" indicate the general character of tacit coordination in following not only agreements, but also any type of rule: if a rule were to spell out all cases of its future application, it would lose its general character and lead to monstrous specifications of rules about the rules for the rules of

> using rules... any rule able to coordinate participants' moves (whether implicit or explicit) *requires a community of players trained in applying the rule*. [emph. add.]

Vollmer believes, as I do, that a "completely explicit," non-fuzzy form of coordination, in which "participants [are] able to give signals that would *unambiguously specify* which moves to make," is—moreso even than impossible—a nonsensical proposition.

Practically, however, utterances vary dramatically in the expected, average variation of inferences across a population and timespan. Policy letter is often designed to minimize the discretion necessary in—and thereby the controversiality of—adjudication. This desire for policy clarity necessarily alters gameplay in turn—most commonly and familiarly by segmenting the legal and illegal at some "bright line" joint of perceptual conspicuity.

2.3. CHEMISTRY AND LAW

Rules—and the surrogate structures of surveillance and management which give them teeth—do not rid players of their guiding, "algorithmic" values, which take environment as input. Rather, they seem to alter the available modes of player expression, to make some outlets available or unavailable. Like the problem in artificial intelligence of the "nearest unblocked strategy," the values and intentions which gave rise to a banned behavior do not disappear upon its banning. Instead, these desires have merely been re-channeled into the "nearest unblocked" action.

In New York and a number of other cities, indoor dining was banned on account of the recent pandemic. Restaurants, scrambling to stay open, began building outdoor seating areas: first surrounded by waist-level plywood walls, a mostly symbolic boundary—then gradually scaled up, with higher walls to block the wind, and roofing for rain. Indoor spaces had been effectively recreated as unzoned "outdoor eating spaces." All the old human desires, shelter from the elements and the street, had remained, and had slowly routed around the new laws, testing its boundaries of enforcement, reinventing old tradition within a legally defensible frame.

Nor, in most cases, does the banned behavior itself disappear entirely—it has not been made impossible, rather, it has had its cost-benefit function altered to make the behavior *more expensive*. It is always physically possible to leave work after lunch, or cheat in poker—it is merely *costly* (to borrow a signaling concept) be caught.[7] Fair play—defined, at minimum, as adherance to letter, but often encompassing social judgments of spirit and ethos—is maintained by prohibiting known violators from entry into future games. To be caught making cheap or degenerate moves is therefore non-ergodic, and structurally analogous to death in a natural selection framework.

7 Costly signals are sometimes thought of only in terms of the up-front cost of manufacturing a sensory display. This becomes particularly misleading in human affairs, where reputation systems (including ongoing surveillance and social sanctioning) make otherwise cheap signals costly (thus enabling roughly honest communication where it would otherwise be impossible). Post-hoc cost, and risk of incurring such costs, is a significant part of the human costly signaling landscape.

We can see this in America's ongoing drug war, and the explosion of research chemicals[8] in the 2000s and 2010s. These chemicals are called analogues because they are "off-by-one"—near-copies that evade the law's letter even as they deliver similar effects, defying its spirit. Laws like the American Analogue Act have been passed in order to blanket-ban such analogues, but determination of what counts as an analogue, and whether such vague prohibitions are Constitutional, have plagued attempts to apply them.[9]

Spirit-based enforcement is more common and powerful in informal games administered by social sanction, than in formal institutional games administered by bureaucratic policy. Still, even within institutional games, bureacuratic discretion

8 Popular analogue sets include dissociatives MXE, MXP, and DXE; 3-MeO-PCP, 3-MeO-PCE, 3-MeO-PCMo, 3-HO-PCP; psychedelics 4-AcO-MET, 4-AcO-MiPT, 4-AcO-DMT, 4-AcO-DIPT, 5-MeO-MiPT, 5-MeO-DIPT, 5-MeO-DMT, and 5-MeO-DALT; opiates U-47700, AH-7921, U-50488, and U-77891; and amphetamines 2FA, 3FA, and 4FA. The complete list is orders of magnitude longer. A similar legal situation has recently cropped up with the introduction of Delta-8 cannabis.

9 *U.S. vs Washam* rested on judicial interpretations of the meaning of chemical "structural similarity." Washam had imported and sold 1,4-Butanediol, a GHB analogue, from Mexico into the United States; he was arrested by an undercover offer after making a five-digit deal for the substance. Expert testimony in favor of the government pointed out architectural similarities between 1,4 and GHB ("both linear compounds containing four carbons") as well as the body's conversion of 1,4 into GHB. Expert testimony in defense of Washam argued that the two chemicals occupy different "functional groups," categories used by chemists to differentiate chemical structures, properties, and reactivity—as well as the argument that MSG, a legal food additive, *also* metabolizes into GHB in the body. From the majority decision: "Washam argues that there is no consensus in the scientific community regarding whether 1,4-Butanediol has a 'substantially similar' chemical structure to GHB under provision (i) of this definition, and thus the definition is unconstitutionally vague as applied."

plays a non-trivial role, and players will often find ways to excise spirit-violators, even if the letter of policy is on the target's side. Strong spirit-adjudication within institutional or legal settings is associated with soft authoritarianism: members of the institution whose authority and discretion in evaluating the spirit of both law and behavior is unchallengeable. Such decisions need not be publicly transparent, consistent, or "fair"—either because they are enacted on small populations, or because the decision-maker does not answer to the public.

2.4. JUDICIAL FORMALISM

But if such authoritarianism is unappealing, overly literalist legal decisions prove equally difficult to stomach. Let us take *United States v. Marshall*, a 1990 7th Circuit Appeals case. The main defendent, Marshall, was subject to a ten-year minimum sentence according to a Congressional law which premised sentencing on the total weight of narcotics sold. A dose of LSD, being just 0.05 milligrams, is typically laid on a sheet of blotter-tabs, or else heavily diluted in another liquid. Thus, as Judge Easterbrook sums up in his majority opinion:

> Marshall's 11,751 doses weighed 113.32 grams; the LSD accounted for only 670.72 mg of this, not enough to activate the five-year mandatory minimum sentence, let alone the ten-year minimum... This disparity between the weight of the pure LSD and the weight of LSD-plus-carrier underlies the defendants' arguments.

Here, we have an extrinsic, three-sided selection game made up of the prosectuion, the defense, and the appeals court. The judge's role is—depending on one's interpretation of

judicial obligation—to comparatively interpret either the letter or spirit of the law against the letter or spirit of the defendant's behavior; this comparison will determine the outcome of the game. Prosecution and defense make efforts to strategically conceptualize both the law and the defendent's behavior in order to alter this process of comparing. The law has been drafted in order to manage the larger wrapping game of American governance, bringing participants into provisional behavioral alignment. This being a nested institutional structure, the judge brings in his own values and desires, which are checked by those selection games in which he plays the role of candidate-object, and the buck only stopping at accountability to the public.

Can we guess, with all our cynicism, the outcome of *Marshall?* The majority ruling concluded that blotters were, by definition, "a mixture or substance containing" LSD, and therefore part of its weight. Judge Posner dissented from this interpretation, since the majority decision let to "results so irrational"—other choice words include "whacko" and "loony"—so as to be unconstitutional. The majority opinion readily acknowledges this:

> If the carrier counts in the weight of the "mixture or substance containing a detectable amount" of LSD, some odd things may happen. Weight in the hands of distributors may exceed that of manufacturers and wholesalers. Big fish then could receive paltry sentences or small fish draconian ones. Someone who sold 19,999 doses of pure LSD (at 0.05 mg per dose) would escape the five-year mandatory minimum... Someone who sold a single hit of LSD dissolved in a tumbler of orange juice could be exposed to a ten-year mandatory minimum. Retailers could fall in or out of the mandatory

> terms depending not on the number of doses but on the medium: sugar cubes weigh more than paper, which weighs more than gelatin. One way to eliminate the possibility of such consequences is to say that the carrier is not a "mixture or substance containing a detectable amount" of the drug. Defendants ask us to do this.

And yet, ultimately it concludes:

> Distributors pick their poison. The penalties are plain for all to see. They decide what drug to peddle, on what medium... The Constitution does not compel Congress to adopt a criminal code with all possibility for unjust variation extirpated. Experience with the guidelines suggests the reverse: Every attempt to make the system of sentences "more rational" carries costs and concealed irrationalities, both loopholes and unanticipated severity.

Moreover:

> extracting LSD from blotter paper and weighing the drug accurately may be difficult. One dose is an exceedingly small quantity of pure LSD... Congress rationally may decide to avoid a costly and imprecise process.

Here we see many of this book's themes play out: the impossibility of completely irradicating loopholes and concealed irrationalities, and the role that practical questions about the cost or ease of measurement (and precise, objective measurement in particular) plays in determining the letter of the law, and by extension, institutional incentive structures. (And by extension, the moves which players deploy within such structures.)

The dissent takes a linguistic tack as well, Posner framing disagreement as a theoretical conflict—with a human life on the line—between "the severely positivistic view that the content of law is exhausted in clear, explicit, and definite enactments by or under express delegation from legislatures" and a more pragmatic view of interpretation, which casts language as servant of spirit. This first, positivistic view is frequently termed formalism or textualism—the idea that "legal problems can be solved in a quasi-mathematical way."[10] "Judge Posner," David Strauss writes, "has never allowed what then-Judge Cardozo called 'the demon of formalism' to 'tempt the intellect with the lure of scientific order.'"[11] As we will see, it is partly the desire for what the philosopher C. Thi Nguyen calls *value clarity*—an objective, cut-and-dry perspective reminiscent of scientism—which drives surrogation, and causes overly literalist spirit-letter problems.

The specific means by which the majority helped reach its decision are frustrating to the linguistics among us. The court used precedent from an earlier case, which after consulting a dictionary came to the conclusion that an LSD blotter fit its definition: "a 'mixture' may... consist of two substances blended together so that the particles of one are diffused among the particles of the other." This literal match was sufficient for their ruling. (And yet, Strauss writes, in another prescriptivist-descriptivist split, this dictionary definition has little to do with normal English use of the word "mixture," which would never call a water-soaked piece of paper a *mixture* of paper and water, or a piece of paper soaked in salt water and dried, with the salt crystals remaining, a *mixture* of paper and salt.) A dictionary definition, like all definitions,

10 "The Anti-Formalist," *University of Chicago Law Review* 2007.

11 Ibid.

is a surrogate which attempts, but inevitably fails, to cover all cases which a native speaker might term a mixture, while excluding all cases a native speaker would not. "A human is a featherless biped," so Diogenes plucked a chicken and held it aloft, triumphantly.[12]

The positivist or "formalist" approach, which we'll contrast with the holistic pragmatic approach, is deeply intertwined with the phenomenon of surrogation; as a literalist approach to language, it mistakes surrogate for surrogated, meaning for messenger.

But it is worth exploring the practical advantages of the formalist approach, which underly the use of institutional and management surrogates more broadly. First, a Constitutional provision or Congressional law is not the product of a single designer, with a single spirit of intent, but rather the result of a dynamic process within a committee of rivals. A bill must pass both chambers of Congress and then the Executive chair; at each stage, there will be voters or drafters with different intentionalities or interpretations of the wording of the law under consideration. Textualism is, in this frame, a pragmatic avoidance of this chaotic, distributed intentionality in favor of their one common denominator: the actual letter of law as written, agreed upon, and passed. How can we meaningfully speculate, in such a system, what Congress "meant" or "intended," when the reality is

12 Analytic philosophy has learned this lesson the hard way—by attempting and failing, repeatedly, to generate definitions which are both "robust" and "elegant," that is, are concise criteria without false positives or negatives when tested against the intuitions of a native speaker. Analytic's 20th C history is marked by failed attempts at factoring terms like "causality" or "morality," each attempt met by counter-examples. [Reason 2020: "Conceptual engineering"; Bishop 1992: "The Possibility of Conceptual Clarity in Philosophy"]

a loosely coordinated kludge which judges must attempt to reverse-engineer decades or even centuries later? We see this desire to avoid questions of intentionality in *U.S. v. Marshall*, where the majority opinion notes that "even laws that resulted from mistakes in the drafting process or ignorance in the halls of Congress survive if a rational basis may be supplied for the result," and cites *U.S. Railroad Retirement Board v. Fritz* and *Delaware Tribal Business Committee v. Weeks* by way of example.

Second, textualism appeals to ideals of public transparency in the tradition of Hammurabi's stele. A population must be able to transparently know the rules of the game they are playing. Oliver Wendell Holmes writes: "We ask, not what this man meant, but what those words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used... We do not inquire what the legislature meant; we ask only what the statutes mean." How laws are understood is, from this perspective, more important than the intent behind its passing—since individuals will act according to the law as understood (and not as intended). Letter laws are here a formalizing reductions which, in their lossy compression, gain the advantage of minimizing vagueness and routinizing decision-making. They are similar to job performance metrics or the grade-point average: their major benefit is removing vagueness and subjectivity from the decision-making processes—but that vagueness has not disappeared from reality, it has merely disappeared from the evaluation, which is suddenly unable to account for it.

Third, a textualist might note that, even if textualist rulings lead to an improper execution of Congressional spirit, a future Congress may always alter the wording of the laws in

order to better represent said spirit. In this way, there is a cybernetic system of alignment between spirit, letter, and extension (i.e. the letter's application). Spirit is translated into letter, which is applied; letter is in some sense a "theory" of the spirit. If the outcomes do not seem to align with spirit, the letter is updated, and so on. Judges, by executing the "program literal" as it is passed down to them, and minimizing whatever normative discretion they can, task policymakers with the continual refinement of their letter specifications. "That Congress could have written better laws does not mean that it had to. Amendments to the criminal code may be in order, but they are not ours to make under the banner of constitutional adjudication."[13] By 1993, Congress passed an amendment specifically altering its guidelines for LSD, in order to prevent the "loony" results of the *Marshall* case.

13 Eastbrook, *U.S. v. Marshall* 1990.

3. Formal Games

We will focus first on the formal selection games which are characteristic of institutional structures. Formal selection is done with respect to a system of law, that is, a system of letter-surrogates according to which evaluation proceeds either deterministically or narratively, using the law as a basis for justification.

Just how much discretion is available, when evaluating play comparatively to some letter, varies. Simple machines and programs will automatically dispense a reward if the programmed criteria are met; bureaucrats can often allow non-trivial leeway, allowing in some greater ability to adhere to spirit, but at the cost of occasional nepotism, corruption, personal bias, etc. Many de facto laws differ from their de jure correspondents—speeding, in the United States, is typically only punished when drivers are going around 10MPH, or perhaps 15-20%, over the posted limit. However, this informal norm of enforcement enables "speed trap towns"—towns, typically positioned on a major interstate, where, in order to bring in local revenue, police officers ticket drivers who exceed the posted limit by even a few miles per hour. King City, California, was infamous in my childhood for being such a town, and those who lived nearby, or drove through it regularly on travel, quickly learned to watch for tell-tale signs of the city limits' approach—a certain patch of trees northbound, a certain gas station southbound. In this way, an interesting emergent effect is informally achieved: the more a driver is local to the area, the less likely they are to be ticketed; there is in effect an ingenius selection game which identifies only tourists and outsiders to subsidize the local economy.

To illustrate internal-game dynamics, we can use a classic (if somewhat tired) example from economic theories of perverse incentives. A city pays a bounty on dead rats (internal game) to improve sanitation and health outcomes (external game). At first, the rat population drops, as citizens are encouraged to clean up the streets. Then, a few enterprising individuals begin breeding rats in their basements, costing the city enormous amounts of money in bounties while having a negligible or even net negative impact on sanitation and health.

We can note a few things about this game by fleshing out its incentive structure or reward function—how it selects players for desirable or undesirable consequences, thereby modulating player behavior. The bounty is the game's *reward*. The rules by which the reward is conditionally dispensed constitute its *letter*. This letter is an attempt by the designer to implement some intended *spirit*: the holistic, inherently vague style of "proper play" the designer intends to incentivize, in order to accomplish a holistic, inherently vague goal in the external game. Finally, there is the game's *metric*: the method for monitoring a player's behaviors, and determining an interpretation of reality which can be measured against the letter (that is, measured by the letter) to selectively dispense the reward. Such a system requires an evaluating agent or *evaluation mechanism* (programmed, judicial, bureaucratic, or otherwise) to interpret the player's accomplishments against the game's spirit-conveyed-in-letter.

In theory, such an incentive structure structure can come apart in a number of ways. First, the style of play the designer wishes to incentivize may not, in fact, accomplish his

desired end-goal. Second, the formalization of the designer's spirit into letter may fail to adequately represent his spirit in all its holistic underspecification; styles of play which he wished not to encourage and reward may dominate the game as it objectively exists in its actual rules, as opposed to as it subjectively exists in the intended spirit of its designer. But most often, the point of failure will not be that the spirit is illegible or overly obscure to players, but that the players willfully neglect the spirit in favor of the letter, or in favor of the "real game," that is, the persuasion of the arbitrating judge to disepnse the rewards by any possible means, which can include dummy rats, or bribes and blackmail. If a rat catcher submits and is compensated for convincing faux rats, and is never caught, he has violated only the symbolic rules of play; in physical reality, he has effectively made a winning move, i.e. one which effectively dispenses the reward.

Thus, playing the game "fairly" is a form of either a lack of imagination, or tacit cooperation—take your pick—that is, a choice to act in a way in which one makes minor sacrifices to one's own strategic advantage—self-handicaps by accepting certain, potentially winning tactics as "out of bounds"—in order to contribute to the greater function of the activity, or the superorganism which has created the activity for its ability to win the superorganism's "external game." (For instance, the military has created a system of medals and awards to incentivize valor to better win wars.) And there are two levels of cooperation—one, adherence to the symbolic rules (letter of the game); two, adherence to the spirit of the game. Many avoid breaking symbolic rules purely from self-interest: there is typically a non-trivial cost if caught. But adherence to the spirit bears less incentive, because any system of evaluation and reward dispensation which wishes to be efficient,

transparent, and objective-seeming cannot punish behavior which conforms to the letter of its law. In some cases however, when game play is publicly visible, the audience can successfuly incentivize spirit adherence among at least *most* players, by imposing a reputational cost to players who win by technicality or through unlawfulness. In professional sports, players who win based on technicalities, or succeed through unconventional play styles, are often accorded less prestige and recognition from audiences or commentators.

A few notes before we move back to rat-breeders:

1. Social, subjective judgment may have the advantage of being able to detect a game's spirit: although they may disagree in some cases, human beings's subjective determinations are often superior to "literalist" interpretations (that one might see in strict judicial formalism, or in computer programming) to identify behavior that is against a game's spirit even as it complies with the game's literal letter. Formalized and "objectivized" decision-making modes lack the context- and intent-sensitivity required for nuanced application of spirit, principally because they are not psychological—like Dionysus, they either lack theory of mind or are unwilling to exercise it out of principle. Furthermore a human evaluator has the ability to "zoom out" and contextualize a ruling in light of the external goals of the game host. Empirical study as to how broad or narrow consensus is on various game spirits remains to be done.

2. The cost that accompanies such play includes secrecy and deception, but also a lack self-esteem, a feeling of imposter syndrome, a feeling of guilt or undeservingness. It is beyond the scope of this text, but our sense of morals, our sense of self-worth, and other (likely culturally conditioned)

psychic needs complicate economically rational play with a notion of emotionally rational play. In some meaningful sense, capital—be it money, status, credentials, etc—which is typically used to ground strategic theories like the one in your hands—is merely a means or instrument to accomplishing higher goals like "life satisfaction," "happiness," or "the good life"—and to reify capital as an ends in itself would be to commit the very surrogation mistake (see §4.4 Fetishizing Means, p. 105).

Now, we can consider the enterprising breeders as "defectors" even if technically speaking, they are abiding by the symbolic rules. Their play style undermines the larger purpose and function of the game itself, even as it advances their own interest. It degenerates the game's telos. This dynamic illustrates the fundamentally adversarial relationship between a wrapping "game" (and its enforcers and designers), and the players who are wrapped inside this game, themselves self-optimizing within the letter of rules outlined by the game designer.

Finally, we can note that the incentive structure of the rat-catching game is a structure of surrogates. First, the letter stands surrogate for spirit. We have learned from Midas to specify edge-cases when asking prankster gods to make our dreams come true. Second, the measurement or metric used to dole out rewards—upon comparison with the letter of dispensation or punishment, via the reactive ritual—is a surrogate. Even if the fully specified letter of the government policy, in hoping to control rat populations, successfully appends "rats caught while running loose, which were not raised by oneself or one's accomplices…," etc etc, there is still the problem of monitoring and observation. It would not be possible to sufficiently surveil a population in order to

ensure that citizens were, in fact, playing by the fully specified game-spirit. So some easily observable surrogate, which somehow correlates or corresponds (logically, statistically, metonymically, etc) has been erected as the real (as opposed to idealistic) basis for doling out rewards. Here, that surrogate is the apparent possession of a dead rat. The failure of the surrogate to stand robust to degenerative play, that is, to be "gamed" by players, is both a failure of surrogate specification (letter standing place for spirit) and surrogate metrics (observable or "manifest" variables standing place for hidden or "latent" variables).

If you are a critical reader, you will have noted that some degree of surrogate metric—observables standing in for non-observables, and being extrapolated in an attempt to create a full portrait of the entity in its non-observable entirety—is present in all human interaction, which gives it its "optikratic" character; appearing is in many cases becomes as good as (functionally equivalent to) "actually" being.[1] *Optikratics*—briefly, the idea that socially evaluated games are ruled by appearances—is a cousin concept of our banal sensory and linguistic once-removals, and perhaps another super-set.

Still, this inevitability and inescapability of surrogates, in both formal and informal games (e.g. dating) does not mean that surrogate systems cannot be more or less perverse, or that internal games cannot be made more or less gameable. It is just to say that amelioration, and not a full "cure" is what is on the table for us, in our daily and institutional capacities as evaluators.

1 The surrogate incentives give way, inevitably, to degenerate play.

3.2. SURROGATE MEASURES

Recall that in statistics, latent variables—variables of research interest which are hidden, nebulous, underspecified, or inaccessible to direct study—are instead measured indirectly, through a manifest or proxy variable, in a process known as operationalization. We begin by understanding one bare-bones form of surrogation that is most analogous to a proxy variable. This "mere" surrogate measure stands in contrast to a surrogate metric, as we will see.

Australian counterinsurgency expert David Kilcullen writes, in "Measuring Progress in Afghanistan," of American military efforts to provide a surrogate measure for progress—as well as the ways such efforts, having chosen over-simplified or crude surrogates, result in a poor understanding of the situation on the ground. SIGACTs—military jargon for "significant activities" such as suicide bombings or insurgent attacks—have long been employed as a surrogate measure for American military progress, with the "assumption that more SIGACTs are bad and fewer SIGACTs are better." This assumption, on scrutiny, quickly breaks down:

> 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.

Thus, the surrogate measure produces a picture that dramatically misunderstands dynamics on the ground by collapsing important distinctions. The correlation between "American military progress" and on-the-ground violence is all over the place; in some regions and conflicts, it may be a

reasonably accurate heuristic; in others it gives exactly the wrong impression.

But we do not yet have the ingredients in place for a surrogate metric, and with it, the emergence of degenerate play.

3.3. SURROGATE METRICS

By "metrics," to be clear, I do not mean specifically quantitative yardsticks—merely yardsticks, or standards of comparison, in general.[2] That is, a surrogate metric is a surrogate measure which is used to preferentially reward measured agents.

Selective reward is critical. Unless selection pressure is exerted on the measured subjects—an *incentive* for them to be evaluated one way vs. another—there is no degeneration of subjects' play (and with it, degeneration of the game or hosting institution). In other words, without selection pressure we do not yet have a full-bodied selection game, because the measurement is of no consequence to the measured. Degeneration requires, at the very least, the introduction of competing agents who are preferentially treated according to their evaluation by the surrogate measure. Selection takes care of the rest: given enough time, agents with play styles who best pass the selection tests will survive. But if these agents are, further, able to discern at least in broad strokes the basis for their evaluations (and by extension, their preferential treatment), then degenerate play will surface sooner rather than later, as agents can consciously and actively scrutinize and exploit weaknesses in the surrogate measure.

2 Furthermore, these yardsticks must be reasonably formalized, such that the game does not function anti-inductively.

While surrogate measures may be poorly chosen, in the banal sense of being poor proxies, they uncouple somewhat randomly, as the normal product of environmental change. Surrogate metrics, on the other hand, are actively and not passively decoupled from what they stand surrogate for.

To fully understand surrogate metrics, and the degenerate, decoupling play it gives rise to, we must establish the adversarial nature of gameplay by way of example.

Consider the "spread"—a dominant strategy in competitive scholastic debate. Debate's rules penalize unaddressed arguments as "dropped" or conceded; as a result, there has been an arms race toward greater and greater verbal speed. Competitors attempt to bring up as many arguments as possible in the limited minutes they are allowed each round; this forces opponents to, with equal speed, address all raised points within their own limited allotment of time (or else effectively cede the round).

To recap: What is on display here is the adversarial relationship not just between players of a game, but between a game (anthropomorphized) and its players. Grounding this conflict in actual human beings, instead of the anthropomorphized "game," we can say that judges, game designers, institutional hosts, etc institute a formal game in order to encourage certain styles or strategies of play; this underspecified intent we have called its spirit. (Such a spirit can be argued to inhabit even evolved or decentrally designed games—more on this later.) While such a spirit is nebulous and difficult to pin down, its existence is testified to by a shared felt sense, among players and observers alike, of cheap play and winning by technicality—indeed, these felt judgments show high degrees of overlap, controlling for the loyalties and interests

of observers.[3] And it is demonstrated in the continual readjustment, by judges and systems administrators, of the literal letter of rules, such that they better reflect spirit and ward off cheap play. These basic dynamics are present in games from Constitutional law to professional sports.

The spirit of the debate game, in some meaningful way, is causally connected to the "point" of play in the first place—the larger, pragmatic purpose that play accomplishes, which can be lofty—simulative education—or base—as in entertainment. These pragmatic functions provide a justification by which judges and administrators alter rules and either prohibit or penalize certain types of play.[4] It is also the social "spirit" which is created and reified (makes itself felt, in players' actual behavior) by discourse around the game, which socially sanctions or encourages styles of play. Scholastic policy debate was established and fostered, throughout the 20th century, in a spirit of civics education—training toward some ideal of public and political discourse. Today, due to "spreading", it is largely unintelligible to uninitiated audiences, who cannot parse debaters' rapidfire speech, let alone the arc of their arguments (which prioritize quantity over quality, a values hierarchy that inverts our usual standards

3 That is, players and their associated "parties" (allies, fans, benefactors, benefactees) who are advantaged by degenerate play are incentivized to argue on behalf of—and, by extension, actually believe—that their play is legitimate and in accordance with the game's spirit. Whereas players and their associated parties which are disadvantaged by degenerate play are obviously incentivized to condemn it.

4 Note the close connection between hosting and refereeing a game, on the one hand, and institutional accounting practices on the other.

of persuasion[5]). Somewhere, the spirit of debate—and with it, its founding function—has been lost to degenerate play.

Players in the debate game, first and foremost, were not just measured through surrogates—they were then subjected to the outcome of that measurement; they were evaluated, and then preferentially treated—that is, selected for wins, losses, and titles—according to the evaluation results.

Next, they were able to gain an awareness of what basis they were being evaluated on; that is, of the surrogates put in place to objectify the evaluation of "quality." In contemporary society, the rulebook of many games is made publicly auditable out of a desire for transparency and fairness (*cf.* the Stele of Hammurabi). But these benefits come with a trade-off: players engaged in an adversarial relationship against the game itself are given an advantage in degenerating the efficacy of—by optimizing toward—the in-place surrogates. When the surrogates used in evaluation are unclear, one cannot very well optimize toward them—the best available strategy is slow adaptation or evolution toward success, preserving tactics which pan out in wins and abandoning those which do not. But evolution is painstaking where the application of abstract intelligence is rapid: when players can study surrogates, they can deductively reason their way to winning strategies, and optimize for those specific traits which will best please the censors (or "gatekeepers"). Drug smugglers, to give a ready example, are closely acquainted with the

5 As an illustration of the idea that it is "surrogates all the way down," consider that persuasiveness is, itself, a surrogate quality standing in for that harder-to-discern quality "correctness." Much has been made, dating back to Greek Sophism and Roman oratory, of this surrogate, and the flawed pedagogy that results from "teaching to the test"—that is, winning over an audience through rhetoric, rather than for being in possession of a superior stance.

technical details and functioning of the systems and tools that screen international shipments at customs. Their packages can then be carefully designed and disguised in order to thwart customs' detection heuristics—for instance, placing contents in packaging that deflects x-rays. But if a new surrogate were put in place—for instance, specifically searching only those packages that contain x-ray-deflecting material, or using dogs' sense of smell—then a player strategy previously optimized would become radically unfit, evolutionarily, in the new system—would become a losing strategy.

With an understanding of the surrogate rules—the letter of the system—debaters were able to identify degenerate tactics such as the spread. Crucially, there is nothing especially reprehensible about degenerate strategies; they are the ordinary condition of a self-interested agent within a competitive incentive system, and need not involve such drastic moral tradeoffs.[6] (Our society, having limited resources, status hierarchies, and relatively exclusive mating arrangements is inevitably competitive in such a way.) Many gamers (incl. David Sirlin) prize the pragmatics of degenerate play, and scorn as "scrubs" those who voluntarily abnegate themselves from such play.

Further, it is not so much that players are "degenerate," but that their play itself tends to degenerate and undermine the original (or, at some level, desired) spirit and function of the game. Indicators that play may be degenerate are found in complaints, by both observers and other game players, that a certain strategy is "cheap." Objections frequently include some acknowledgment that the play is "technically" legal—that is aligns with the game's letter—but is, nonetheless, a

6 "Hate the player, not the game," in folk parlance.

cousin to cheating. (Cheating being behavior that violates not just the spirit of a game but its letter; where a letter-abiding judiciary is limited in prosecuting spirit violations, it does much better with prosecuting letter violations.) In this case, the spread is "degenerate" insofar as it goes against the founding civics-oriented spirit of scholastic debate.

We can also revisit Kilcullen's example, in which the military, attempting to measure some abstract and underspecified "progress," instituted as surrogate the rate of violence, or number of SIGACTs, across regions. Recall that this metric obscured, by over-compressing, a situation on the ground in which low-violence areas were just as likely to be enemy-controlled as American-controlled. Now we can introduce a variant of the situation—necessarily simplified, but still illustrating real dynamics—in which, first, the military gives greater attention to high-violence areas (ceding low-violence areas as completed goals), and second, the Afghan resistance, by intercepting American military intelligence briefings, has become aware of the military's system for evaluating and attending to different regions. Here we have a picture where all the criteria of a surrogate metric—that is, all the criteria of a full-bodied selection game—and not just surrogate measure, are in place; we should expect the surrogate to degenerate not merely by chance environmental drift but forceably, through degenerate play. There is an an evaluating body whose behavior has consequences for evaluated players—that is, the evaluative system results in preferential selection or asymmetrical outcomes for players on the basis of the evaluation. And there is knowledge, by evaluated players, of the basis for this evaluation and, by extension, their preferential outcomes. At this point, a fairly predictable set of strategic behaviors emerge, with the Afghan fighters

attempting to redirect American attention and efforts away from regions the fighters find strategically valuable and toward regions the fighters find strategically irrelevant. And indeed, as Ben Connable shows in *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*, Vietnamese insurgents did often refrain from violence in order to avoid US military detection and maintain "freedom of movement." The surrogate metric is far worse than random—in such a situation, it can become *negatively* correlated with the target it had hoped to stand in for, after being forcibly uncoupled by strategic agents manipulating the dataset.

3.4. DECISION RULES & MAGIC WORDS

In some formal selection games, evaluators—being themselves deeply nested in a strict oversight system—are forced into simple, quasi stimulus-response patterns of action corresponding to their evaluative "inputs" (the expressive outputs of the evaluated agent). We can think of this bound behavior as a *decision rule*: if this, then that. As soon as an evaluated party "checks a box" or presents in a certain, formally described way, the evaluator must automatically—that is, with a minimal of holistic, contextual judgment—begin a certain response protocol.[7]

One way that the behavior which triggers institutional decision rules is sometimes described is as "magic words." As Freddie DeBoer reflects of the psychiatric system, "magic

7 Of course, psychiatrists are not simpletons, and the "magic words" are not automatic admissions; there *is* some context-sensitivity in involuntary commitments. However, because of the *structure* of lawsuit liability, a patient who even jokingly or rotely asserts that he is a "danger to himself an others" risks a potential lawsuit for the practitioner should he attempt suicide shortly after.

words are a card you can play, a big one." (Note the games language; the importation of idioms from gambling and sports should serve as linguistic evidence for the game-like quality of life.)

> The magic words, of course, are some version of "I believe I am a danger to myself and to others." It's a conversation starter; you can get the most bored attending psychiatrist's attention that way. If you say such things and they let you go and you stab somebody, they can get sued, which they obviously don't want. And the law gives them the tools to prevent that. In most places in the country, that kind of talk can get you involuntarily committed.[8]

An institutionally competent member of in-patient facilities, Eric Berne writes, can

> choose at will between (1) staying out indefinitely (or as long as the family finances hold out), (2) being transferred to a less demanding environment such as a state hospital, or (3) going home whenever he is ready. He also learns how to behave in order to be readmitted.[9]

Berne may be overstating the case, or overestimating the number of competent patients at psychiatric facilities, but

8 DeBoer, "When You Have Come Apart," 2021. Similar schemes have been shared on social media for e.g. gaming a GP's attention by emphasizing the impact of symptoms on a patient's ability to perform Activities of Daily Living (ADL), exaggerating impact, and generally performing what Goffman would call "dramatic realization": "Super over-do your ADL: if you can't put on socks, walk in barefoot. [Doctors] need to see evidence of how your problem negatively affects your life when you walk in the room" (@maiab, Twitter 2022).

9 Berne, *What Do You Say After You Say Hello*.

this general principle of simple stimulus-response behavior by an institution opens it up to being "gamed," manipulated much like one would a machine. And indeed, the difference between what we typically think of as a machine, and a full-bodied agent, *is* their complexity. The machine is simpler, more deterministic, and more predictable; one can quickly learn which inputs yield which outputs, and then tailor behavior accordingly. And the machine, in its behavioral rigidity, cannot adapt to being "played"—it cannot dynamically update its "algorithm" or operating procedure. This makes it, technically speaking, "stupid." This difference is one that will be repeated in following chapters—the timescale by which a system can adapt to new play styles. Evolution is "stupid"; people are "smart."[10] The law is stupid, fashion is smart. And so on.

The missing step, of course, is a more complex, contextual mode of interpretation. Kilcullen again, "Measuring Progress In Afghanistan":

> 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

10 This is not to say that testing, natural selection, etc are less *reliable* designers than rationality, inference etc—indeed, given a certain level of environmental stability, it is a *more* reliable designer, and arguably *more* context-sensitive. The problem is that, when it is "up against" actors who operate at a faster timescale, those actors are incentivized to increase the level of environmental instability (turnover) in order to gain an advantage.

> detect trends... These trends may not be obvious to personnel who are on short-duration tours in the country.

A similar stupidity is found in the surrogation of individual cases to the category which the cases belong to. In the compression process of accounting that parallels institutional ritualization of decision-making, various items are grouped by analogy, typically according to or measured on a single axis, and then each unit is treated as fungible, or functionally identical. An employee is an employee, a homocide is a homocide, obscuring meaningful, project-relevant differences, differences which might make all the difference.[11] This blindness to indexicality and individual variance is part of surrogation blindness.

3.5. COMPETING AGAINST LIARS

When systems are set up this way, so that they are "stupid" in the sense of context insensitivity, surrogation blindness, and low dynamism, most individuals who are evaluated by them adapt. Competent interviewees, in a psychiatric or child custody evaluation, will not mention that they "on occasion" polish off a wine bottle in an evening by themselves, even if the occasional inebriation causes no real harm to the child, and even if they are far below the average level of alcohol consumption.[12] A parent under a custody evaluation

11 Dan Luu, "Individuals matter."

12 DeBoer uses the term "checklisting," which like "magic words" captures a certain procedural simplicity, in which displaying a set of expressions can semi-automatically lead to a procedural outcome:

> When the doctor asks you about drugs and alcohol, what do you say? If you're sure that those aren't the real problem, you may want to say that you never

will not bring up that they "in college" took cocaine, even if experimenting with hard recreational drugs is relatively normal even among well-adjusted, competent adult members of society. They will not, in a psych eval, say, "Sometimes I have suicidal thoughts, but they're not *serious*, and doesn't everybody fantasize about killing themselves at least occasionally?" Nor will they admit to ever having "hallucinations," although minor visual and auditory hallucinations are ubiquitous. They will not, in an adoption application, check the box which says they have struggled with depression in the

---

do drugs and you barely drink. "Couple drinks, couple times a month, with friends from work, that's it," might be what you say if you drink a bottle of wine a night. Chalking up your problems to substance abuse makes things nice and comprehensible for the people diagnosing you and there's programs they can sign you up for and the next thing you know it's six weeks later and you're sitting in an AA meeting while your untreated schizoaffective disorder rages inside of you. I'm not saying that I know better than the doctors, or that you will. I am saying that you need to make sure that you don't get checklisted as an addict if that's not your real problem.

DeBoer further recommends emotional display as a way to be taken seriously by institutional evaluators:

[D]on't calm down. Your panic and emotional devastation are your most valuable tools when you're trying to get a cold and indifferent system to give a shit about you. The people in that system know, in some remote sense, that someone can appear relatively calm and be in deep need of psychiatric care. But when you're trying to wring that care out of them in a busy ER on a Friday night? They see you looking minimally composed and think that all you need is a cup of tea and a good talking to. That's why, when people contact me in the throes of a [mental] breakdown and ask to talk about going in—which happens more often than you'd think—and they ask me if they should take something, I always tell them no. No Xanax, no Benadryl, no glass of whiskey. Nothing that will artificially restrict the natural expression of your illness. Because it's only that expression that can compel the lawsuit-avoidant edifice of emergency psychiatric medicine to care enough about you to get you into treatment.

past, even if the vast majority of adults struggle with depression and make perfectly fine parents.

Rather, they will say, "I do not drink more than one glass a night" to represent that their drinking is not a problem. They will say, "I do not struggle with depression," to represent that their depression would not be a child-rearing problem. They will say "I never think about suicide," "I never have hallucinations," and "I have never taken drugs." And in their minds—and to some extent, in the evaluation game itself—this will be equivalent to being honest to the *spirit* of the questioning even as it is literally dishonest.

The issue, of course, is that those who are immaculately honest, or treat such questionings as "literalists," must compete against those who answer the "spirit" of the evaluation instead of its "letter." An alternate framing might be that the evaluation system is literally broken, as a result of lawsuit liability or dogmatic rigidity on the part of the overseers, and that it can only be made to work through systemic deception. In this way, unfortunately, whose who deceive within the system in some sense "defect" on fellow players: were the system to deal with honest literalists only, it would quickly collapse—there would be no acceptable applicants for adoption, no patients mentally sound enough to forego involuntary commitment, no parents reliable enough to maintain custody—and be forced to abandon its rigidity. Instead, its rigidity is able to continue because of the "participation" of deceivers in its structural absurdity.

There is reason to believe that, even in informal games of evaluation, evaluators depend on categories in order to make decisions. Just as the doctor must put the patient in a "box" to determine a treatment plan, our perceptual schemas are

deeply "typified" (to use a term from Alfred Schutz) in that we use perceptual cues to determine whether an object is, e.g., a dog, or a pitbull, or a labrador—and then, once we have typified the perceptual object, we activate a behavioral or interactional protocol—which may be more guarded or defensive, in the case of the pitbull, or more friendly in the case of the laborador. But more on this in the chapter on informal games.

3.6. THE GLOBAL KNOWLEDGE GAME

Surrogation is both crucial and corrosive to the function not only of coordinated superorganisms (such as the military, police department, and legal system) but also to stigmergic distributed human projects such as what Sarah Perry dubs the "global knowledge game"[13]—the ongoing attempt to discover global, non-indexical truths that can be reliably used as the basis for prediction and engineering.

There are numerous surrogative issues 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 in the social sciences) discussed with respect to superorganisms. Moreover, surrogation is common across knowledge-oriented fields, such as education, where we've seen controversies over (e.g. the spirit-betraying) "teaching to the test," or more

13 Literal Banana 2020.

blatantly rule-violating actions such as teachers manually altering students' Scantron responses.

Additionally, in the "inexact sciences"—that is, those fields which are attempting to mature past their qualitative roots and into a more quantitative or empirical science, for instance psychology's rejection of phenomenological or psychoanalystic methods in favor of lab research—there is a problem of wanting to "grow up already."[14] 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. I'll use Tal Yarkoni's critique of social psychology, "The generalizability crisis" (2019), to understand the sociological motivations that lead GKG institutions and players into surrogated divergence.

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 notes, thinkers across the inexact sciences[15] have raised the alarm on such issues for decades, in some cases for upwards of half a century—but the paper is a valuable work of information logistics insofar as it compiles and makes sense of the linguistic, inferential, and surrogative problem inexact fields face.

14 See Reason 2021, "Notes on the Inexact Sciences," for discussion.

15 E.g. Gerd Gigerenzer, Paul Meehry. See also Gigerenzer's writing on the "surrogate idol" of a universal method.

80 First, a psychological construct, in order to gather evidence as to its "existence" or "nonexistence"—and even here there is a whiff of philosophical 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 precision. 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 finding depends, 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 finding is re-translated back into language via generalization or loose induction. The coarse metrics disappear as we re-enter the realm of descriptive 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.

3.7. EXAMPLES OF DEGENERATE PLAY

Before advancing to a discussion of surrogation in informal games, we can look at several examples of degenerate play in action, to better understand its dynamics.

Flopping in athletics

In modern limited-contact sports, most prominently professional soccer and basketball, a system of officiating is in place with the goal of reducing dangerous contact (or, if you would rather, of reducing injury), and thereby of allowing enough physical space between players that the game does not devolve into a tackle sport. (Were tackling not specifically prohibited, we could imagine basketball quickly becoming an unwatchable and incredibly dangerous sport.) That is, player safety and audience entertainment are some of the higher-level goals that inform these games' spirits of fair play, and determine the letter of law which is written into officiating rulebooks.

Much like the law, athletic officiating is performed by human evaluators—referees—who reconcile their interpretations of game events against their interpretations of a game's rules. In both domains, player intent, and causal precedence, underly decision-making. As an example of this interest in

causal precedence and intentionality, note that when contact between two players occurs, it is—roughly speaking; there are exceptions depending on sport and circumstance—the player who initiates contact who is penalized.

And although there is significantly less delay, and significantly more transparency, between the event and ruling in athletic officiating than there is in our legal system, there is, nonetheless, a similar high degree of unknowability with regard to the issue at hand, be it a homicide or officiated contact between players. Adjudicating officials must inferentially recreate a historical event based on minimal clues. In professional sports, play is rapid, and there are typically just a few referees on the field of play who have been tasked with monitoring the physical movements (and inferring from them the psychological intents) of players.

As a result, surrogate metrics are implemented by referees; most crudely, and founded on the Newtonian principle that every action has a reaction, we see the heuristic that the player who is most physically impacted or displaced, in the fallout of contact, is the "victim" or recipient of that contact, instead of its initiator. As a result, a phenomenon known as flopping has emerged in these sports, with players "acting out" dramatic falls, head snaps, and injured reactions in order to alter the interpretations of referees. Because this behavior is widely understood, by players and audiences alike, to violate the game's spirit of fairness, the NBA, and many international soccer organizations, have made efforts to combat flopping by penalizing it. But the difficulty of interpreting player intent, or discerning the "truth" behind appearances, has undermined these efforts, and the practice remains ubiquitous in many limited-contact sports.

Affirmative action

Be it in awarding contracts or doling out business licenses, federal and state governments in the U.S. have prominently advertised preferential treatment for organizations owned by women or minorities. The legal fact of ownership stands surrogate for the meaningful sense of ownership, with predictable results: Many male-run government contractors will legally put their businesses in their wives' names in order to reduce the disadvantage they face. Similarly, reparative justice efforts to encourage black entrepreneurship in the cannabis industry, by preferentially awarding licenses, has resulted in many honorary[16] black owners or co-owners, who are paid some small percentage of profits in order to act as a front for white-owned dispensaries. Whatever initial goal the government may have hoped to advance through such programs has been thwarted by the adversarial, degenerate play of the evaluated agents (business owners). That it has happened so quickly since the announcement of these programs is in large part a result of public knowledge of the surrogate metrics, and by extension, of the basis for preferential treatment. We can imagine a situation in which a more black-box evaluative process would stay robust to degenerate play longer—with the cost, of course, that fewer minority businesspersons would be aware of the program, and thereby not especially incentivized to apply for government contracts in the first place.

Similar situations exist in affirmative action programs in universities. One common problem is that, in establishing

16 *Honorary* is another interesting non-technical term in *surrogation*'s vicinity, in that it distinguishes one type of member (or title, or role) from another type which is seen as more "substantial" or "real."

quotas strictly on racial grounds, universities which may have wished to admit disprivileged black American youth have led to the admitting of vast numbers of highly privileged or wealthy black international students. This is not to weigh in, politically, on what "ought" to be the case—merely to note that what was purportedly sought or intended by these institutions has been contradicted by what has in reality occurred; at Harvard in 2004, Henry Louis Gates argued that around two-thirds of the black student body was not of slave descent. If we are being cynical we might emphasize *purportedly* in the sentence previous: taking the "layered" or "nested" institutional perspective on selection games, we might say that the universities were less interested in actual socioeconomic justice, or in affirmative action as a form of reparation for American slavery, and primarily interested in increasing their statistical diversity for recruiting and public relations purposes. Here, the higher tuitions paid by international students, and the global influence that educating elite international students bring, goes hand-in-hand with statistical diversity to furnish their recruiting brochures. In a system of surrogates, it is not "cargoculting" (see §4.4) or "irrational" for an institution to optimize for appearances, since it is on appearances that the institution is evaluated.[17]

This alone is not an example of surrogate metrics or degenerate play—it is merely a poorly chosen surrogate measure—resulting, no doubt, because the evaluative systems had not fully thought-through the actual intervention they wished to enact. Like Midas, wishing that all he touches turn to gold,

17 This point also challenges the standard view of narcissism (excessive concern for image) as pathological or maladaptive; rather, the issue is that narcissists are, in the long-term, in fact poor image-manipulators; they thrive in the relative anonymity and short-termism of modern professional and social life, jumping between jobs and communities freely as compensation.

the spirit of the request is inadequately translated into letter. (Underspecification as a contributor to surrogation problems will be explored in following sections.)

However, there is the closely linked situation in which white students have claimed minority status through some obscure ancestral line—two prominent recent cases are those of presidential candidate Elizabeth Warren, who claimed American Indian heritage in her application to Harvard, and the professor (and previous NAACP chapter-head) Rachel Dolezal, who identified as trans-racial. Were preferential treatment to minority status unknown among student applicants, we can imagine that such disclosures would be unlikely—the applicant is often unaware or distantly aware of their heritage, even if the heritage claim is legitimate—but since it is common knowledge that minority status gives a sizeable advantage in college applications, such disclosures are not unlikely but regular. Any biracial applicant to an institutional entrance game, be it a competitive prep school, university, job, title, etc, is aware of the strategic valence to self-identifying as e.g. white Caucasian vs. Southeast Asian. Again, the adversarial, ontologically tangled relationship between players and game designers—well familiar to Dungeonmasters, lawyers, and parents of young children—is on full display.

Pretty Woman (1990)

The most famous scene from the film *Pretty Woman* takes place at a high-end clothing store on Rodeo Drive, Los Angeles. The shop's sales clerk has a system of evaluation which helps her effectively identify clients based on their financial assets and spending potential—and to then selectively cater to these based on this assessment, which functions to maximize her own own commission. (This commission is the

larger goal of the system, which the evaluation is instrumental to accomplishing.) This clerk cannot possibly know the real spending potential or desire of any customer who enters her shop, but she has limited attention and time, and so she uses surrogates such as their dress and mannerisms in order to make educated guesses and allocate that time accordingly. Inevitably, where these surrogate metrics diverge from reality, she faces (like any other evaluating entity) the possibility of false positives and false negatives: someone who lacks the capital to spend but appears to have it, or someone who despite possessing the capital to spend (and the desire to do so) is not positively identified as such. In the film, Julia Roberts's character registers as false negative, and she is turned away from the store on the basis of her attire—is, accordingly, not allocated any of the clerk's time or attention, nor the resources of the shop, such as the ability to try on garments.

This example helps highlights a dynamic present in many, if not all, surrogative behaviors: the evaluating entity has limited resources—at the very least, the resource of time—and, combined with other barriers to knowing the "true" nature of things—full knowledge is always physically impossible—leads this entity to an economic surrogate. But again, this situation on its own is simply a surrogate measure. But, given that there are agents who desire the clerk's attention, or to try on the shop's clothing, despite lacking the financial resources to "properly" earn it—and given that class markers are common surrogates in high-end establishments, we might imagine players who knowingly rent or steal an outfit worth of high-end clothing specifically to fool such a shopkeeper.

ꝏ

Many of these examples, perhaps most explicitly that of flopping and *Pretty Woman*, land us with a considerable problem in carrying on with this conceptual project. As soon as we move beyond institutional surrogates and quantified metrics, the behavior displayed becomes ubiquitous to human life—we are constantly acting as if, or bluffing our way, or dressing up to impress—and, on the flip-side, judging by proxies, inferring wholes from metonymical parts—and this gets us into nebulous, murky conceptual waters, where surrogation and degenerate play seem to underly all human social life. A world where we live exclusively among surrogates, through surrogates, and for surrogates. The next chapter will explore the conceptual boundaries of surrogation. I believe surrogation is a transitional concept much like Austin's concept of the "performative." One gets the feeling, reading *How To Do Things With Words*, that very much or all of language operates, in some sense, like naming a boat or saying "I do"—as interventions in *social reality*, a means for creating impressions in people's heads about the state of the world.[18] The performative is less a kind of phenomenon, and more a set of examples which especially unambiguously operate off a deeper principle or dimension in all language. As a concept, it participates in a broken ontology—but points the way to a better, reformed understanding of human systems and behavior. I believe surrogation was, and may still be, such a transitional concept.

18 That all communication is manipulation, and some manipulation is mutually advantageous.

4. Informal Games

4.1. NATURAL BOUNDARIES

Nietzsche, "On Truth and Lies in a Nonmoral Sense":

> The thing-in-itself… is something quite incomprehensible to the creator of language, and something not in the least worth striving for. This creator only designates the relations of things to men, and for expressing these relations he lays hold of the boldest metaphors… It is this way with all of us… We believe we know something about the things themselves when we speak of trees, colors, snow, and flowers; and yet we possess nothing but metaphors.[1]

What is becoming apparent is that a system of surrogates amounts to a perspective on the world, a way of seeing. This broadening of scope threatens to obliterate *surrogation*, transforming it from a useful concept to the vague atmospherics of a hand-wave.

Briefly, I wish to discuss the far outfields of this concept, or family of concepts, *surrogation*. I want to identify what is domestic versus foreign territory, and to draw a very fuzzy, gradient boundary between surrogation and these outside lands. But this analogy mischaracterizes the situation—the

1 Nietzsche, in classic torque epistemology fashion, overstates the case; the text continues, "metaphors which correspond in no way to the original entities." But there are necessary, tight correspondences between our map and territory, else our representations would fail in an evolutionary epistemology sense; see following pages' treatment of Popper.

issue, really, is that one patch of the family is nested inside one larger conceptual set, and the next patch a subset of a different wrapping concept, and so on. The concept of diaspora is more illustrative—not so much in its implication of a shared genealogy and dispersion, but in its sense of related subpopulations or subcultures, scattered and embedded in larger super-sets of otherness.

The first super-set of note—a super-set insofar as it furnishes *necessary but insufficient qualities* of its surrogation subset—is Western philosophy's undead subject, and veritable obsession of the Enlightenment tradition. That is: our once-removal from reality, gestured at in the Nietzsche passage which opens this chapter—itself a reference to Kant's *Ding an sich* but an idea which dates to Plato's cave, likely earlier. Our signs are stand-ins for the aspects they pick out; we treat them, in cognitive shorthand, as if they were reality—reify our concepts as objects, are surprised when words break down on us. Our words are referents not just once- but twice-removed from the world, a surrogate for our organized perceptions, themselves representational of the origins of senses.[2] "The map is not the territory," as Korzybski was fond of saying; similarly the symbol is not the substance, and the surrogate not the thing surrogated. Thus we question whether our ontologies are "real" or "fake," and the extent to which our sensory impressions can be trusted as representations of the real. Perhaps the thinker who has made the most philosophical progress on this question is Karl Popper, with his concept of evolutionary epistemology: If there were not a real and deep

2 Importantly, our ability to intervene on the world using formal logics is bottlenecked by the quality of our representational schemas, or ontologies. David Chapman of *Meaningness* and Collin Lysford of *Desystemize* have made compelling cases for the advancement of science as a history of ontological progress.

relationship between a monkey's spatial perception of the tree branches, and their real spatial positioning, the monkey would tumble to its death. But our eyesight also works by proxies and heuristics, and certain contexts can fool it, as in optical illusions. Our visual perception is a tight surrogate to real spatial relationships, optimized over the set of visual experiences that we can be expected with some regularity to undergo, but inevitably, as all heuristics do, failing at certain edgecases or in certain paradigm-breaking situations. (And perhaps breaking down entirely in the face of radical environmental drift.)

> We say the map is different from the territory. But what is the territory? Operationally, somebody went out with a retina or a measuring stick and made representations which we then put upon paper. What is on the paper map is a representation of what was in the retinal representation of the man who made the map, and as you push the question back, what you find is an infinite regress, an infinite series of maps. The territory never gets in at all. The territory is *Ding an sich* and you can't do anything with it… The mental world is only maps of maps of maps, ad infinitum. All "phenomena" are literally "appearances."
>
> Bateson, "Form, Substance, Difference"

The second boundary was pointed out to me by my friend and colleague (the pseudonymous "Crispy") in his essay "Wireheading as Teleological Misnomer." In a similar vein of cognitive shorthand—or proxying and inference—we fall into error by uncritically assuming that a system's name,

originary intention, or public description are interchangeable with its the system's function. (*Optikratics*.) "Names trick you into bottoming out your level of inquiry," he writes. Just because I program an algorithm named "doubleTheInput" does not mean that my algorithm will double the input. No, the system's functionality is neither how it is named and described, nor the intention of its designer (though that intent is the actual function's genesis), nor how it is socially perceived (though the actual function is the genesis for *that* perception). "The problem is that names are generally teleological: a can opener is meant to open cans." Individuals who view the world with an "object-oriented" lens,[3] more than a "functional" lens, often struggle to find functional substitutes for a missing ingredient, material, or tool. Duct tape *just is* duct tape; a recipe that calls for butter requires butter: the pragmatic properties, having been erased by their nominal representative, cannot be found in their functional equivalents; the only operation left to the reifier is an identity check. To the functionalist, butter is a set of properties which it varyingly shares with a host of other food products (oils, yogurts, fatty fruits) any of which, depending on context, can serve as functional replacements.

There are specific cybernetic ways that perceptions, intents, and names act directly on a system to align its real functionality to their image. When a system is intelligent enough to pick up on name, intent, designer, and adapt itself to them, there may be a gravitational pull. (Those American Indian tribes who believed in nominative determinism no doubt saw it played out.) The concepts *hyperstition*, *meme magic*, and *Tinkerbell effects* help map this space, as does the popular

3 In the programming, computer science sense, which shares a name—but not much else—with the 21st C Heideggerian philosophy.

phrase "fake it til you make it": in a world of social proxies and deferrals of judgment, appearances make themselves felt and real. But it remains the dominant frame—that is, the dominant explanatory thesis to which hyperstitional effects stand as notable contradiction—that these genealogical and representational cousins of the thing itself (intent, description, name, perception) are not, finally, equivalent or interchangeable with the thing itself, with the objective functioning of the system—and ought not be reified as if they were.

(And yet we see these mistakes constantly. We confuse good intent with good outcome, or use intent as surrogate for outcome. We confuse the self-representation of an organization with a neutral description of its actual operation, when a name is a surrogate put under tremendous selection pressure, and thereby strategically designed.[4])

The third boundary is, I think, the social version of the first. Going out on a limb, I might speculate that this social version is in fact the original impulse, a soft interpersonal paranoia which underlies our metaphysical and sensory skepticisms. And that is the larger problem, gestured at in the final section of the previous chapter, on *Pretty Woman*, that our entire social life is navigated through surrogates.

What are the qualities we prize most, in selecting a partner, platonic or romantic? Loyalty, care, counsel, company. A professional partner, a colleague or employee or employer? Competence, efficacy, fairness, reasonableness. Of criminals

4 In the abortion debate see the *pro-life* v. *pro-choice* framing. We can call similar techniques "strategic conceptualization": insofar as many issues or decisions (personal and communal) are settled on the basis of verbal descriptions, these verbal descriptions are manipulated, and the specific categories or types that a situation is ascribed, with considerable intentionality.

in trial? Intent, remorse, state of mind, causal role and influence. Of a teacher? Dedication, theory of mind, patience. None of these are directly observable. They are not physical objects, or "properties." They are patterns of behavior; they are designators which represent a set of predictions about how the individual will behave in a variety of circumstances. They are judged by gut feelings, surrogates, and a sample of incidents in which behavior is interpreted as testimony. And yet we must make rapid assessments of trustworthiness, honest, honor: in a used car salesman, in a stranger on an empty late-night street, in a new business or romantic partner. (We partially solve this by withholding, or gradating, serious commitment over iterated encounters. One is not given sensitive state secrets during orientation week, as they say. Access to inner sanctums, architectural, sexual, or otherwise, is often granted only after a long trial of intimate, proximate assessment. For further discussion, see §6.5. Close & Distant Evaluation, p. 190.) We are wary for a new partner defecting on us, "cashing out"; we have nightmares in which some intimate turns out to be "undercover," to be exploiting us. (This is the premise of films like *House of Games* and *Basic Instinct*.)

In statistical language, we operationalize latent variables that we care about, with proxy variables that statistically coincide with the latent variables. In biology, these proxy variables are called signals.

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.[5] Put in economic terms, there is a vast deal of private information which is directly inaccessible to us—there is no way inside the other's mind[6]—and yet which remains salient to our own goals—for instance, another entity's intentions, abilities, beliefs, etc. Instead, we use publicly available information—facial expressions, behavioral patterns, costume—as expressive metonyms from which we can make educated guesses about the "hidden algorithm" that governs others' behavior. Ultimately, insofar as we are interested in understanding entities so that we can predict, and thus optimize around, them, these "originating algorithms" are what interest us more than the specifics of how they are expressed. The facial expression is not of interest in itself; it is of interest because of what it might *mean*—mean in terms of the entity's inner state, and the ramifications of that inner state for the present or future interactions.[7] An individual's conscious self-perception of intent is useful (to him and to us) only as it is a surrogate to his actual future behavior.[8]

5 See William James, *Principles of Psychology* on brute association.

6 And one's mind is opaque even to oneself, for both strategic reasons—see the work of Robert Trivers—and as a function of wireheading—see Freud on repression.

7 For arguments framing "meaning" as entailment to the interpreting subject, see the Pfeilstorch letter series "Meaning of Meaning" (theinexactsciences.github.io).

8 And of course, we can only know his conscious self-perception through the surrogate of his public self-representations—linguistic-explicit, expressive-implicit, or otherwise.

Finally, because our ability to perceive and model and act efficaciously in the world is premised on our surrogate systems, it would be misleading to think of the "surrogation problem" without thinking of its precondition—the surrogation miracle. Surrogation is a capacity, a tool, a tactic. A surrogate is a heuristic which, like all heuristics, functions better over certain problem spaces and probability distributions. A system of heuristics constitutes a framework or perspective on the world, a powerful orientation towards one's environment which—like all powers—is also limited. It is more precise to say that there are problematic approaches and attitudes toward surrogation.

The surrogates we use to read and write to each other form a structure of knowledge and a theory of being. We perform computation less on reality than on these surrogate structures of belief. They form the channels and media available for communication; in so structuring, they also inevitably constrain the space of possible expression and interpretation. Constraint and empowerment are simultaneous, coterminous phenomena. It is tempting to take the cynical view and emphasize our limits; many sound thinkers have. But this is half a story.

Language—as a surrogate system, a representation system, a system of heuristics—underlies much of our reasoning and by extension our acting (particularly our acting-together-in-the-world). Its surrogates are the lines by which our rules are applied or disregarded, the specifications which help guide and focus our intuitive observations, predictions, and interpretations. Our treaties and contracts, at every level of agreement, are written in them. Just as it is true that so many other forms of treaty and contract—so many other "modern worlds"—are precluded, so it is equally true that a

world—our world—is enabled. However lossy our compressions, there is no intelligence without compression. However failure-prone our systems of monitoring or surveillance, there is no efficacious acting on the world without such monitoring.

4.2. INFORMAL SELECTION

Informal selection games lack a "letter" in the sense that formal games are bound together by letter. Some individuals may erect personal laws which they strictishly adhere to, and their behavior becomes functionally formal, but most do not, or do not most of the time. In any case, typifying perception[9] leads us to "round" novel stimuli up or down to the nearest pre-existing categories, which serve as built-in "decision rules"[10] and predictive schemas for interaction. Whether the selection is carried out by an individual or institution (that is, a committee of individuals bound by common law) is less important than the larger trends and consequences that result from having to obey, at least loosely, some letter specification in decision-making, or being able to loosely note and leverage associative trends.

Sometimes, informal selection is undertaken precisely in the attempt to preserve spirit. As in the supervised reinforcement learning techniques popular in machine learning, new selectors are trained (often by their own respective selectors) to develop a "feel" for desired vs. undesired candidates, honing their ability to read and discriminate vibes. Tim Latterner,

9 See Alfred Schutz's social phenomenology.

10 In social studies, these are sometimes called "scripts"; see decision theory as well as Natural Hazard 2021, "Arguing Definitions As Arguing Decisions."

in his *GQ* profile of nightclub owner Paul Sevigny, writes of Paul's Casablanca:

> the man at the door of the club was trained by Sevigny himself. "I really wanted to make a point that there's a different way to do this," he says of bringing on gatekeeper Ludwig Persik. "I was trying to make sure we were on the same wavelength. It's important for me to have people that speak the same language I do and we have the right kind of people. So Ludwig and I would sit in the front window of the Dean and Deluca on Broadway and as people would pass by I would ask him who he would or wouldn't let in the door of the club." Stepping outside, Ludwig corroborates the story of being hired. "It was the end of August and it was really warm out, and he was wearing a Paul Smith suit, even though it was so hot," Persik recalls. "We were sitting in the window and he asked, 'Where do you think they're from?' about a group of people. I said Murray Hill. He laughed and followed up if I'd let them into the club. When I said no, he said 'Good, good, good.' We did that for an hour or two with different people passing by.

Informal games are more anti-inductive, and display regular sociological patterns of solutions—and their decaying value—known as fads. (More in "Evolutions.")

Indeed, it is precisely the assumed adaptive nature of our schemas, which interpret incoming sense data, and determine appropriate actions based on categorical diagnosis of the source of such sense data, which causes us, when such an adaptive nature is lacking, to diagnose a schema's functioning as pathological. Frequently, this lack of adaptive fitness is termed "trauma." In a 2019 discussion of Karen Horney's

Neurosis & Human Growth, I described this general concept as follows:

> an improper environment in childhood causes a deep, underlying anxiety (or feeling of precarity) which leads the child to seek anxiolytic and palliative coping strategies at the cost of real growth. We can call this development non-acute trauma, referring to the banal way an environment routinely shapes one's priors about self and society, such that when one leaves the conditioning environment, previously adaptive strategies become suddenly maladaptive. In an extreme case, and ancient archetype, the soldier returns home, bringing with him an adaptive jumpiness which while useful on tour, causes him to hear gunshots in slammed doors and backfiring engines. We can look back to Euripides' *Herakles* for a portrait: Herakles comes home and, perception befogged by madness, mistakes his children for enemies, slaying them with poisoned arrows.

And I quoted my Pfeilstorch collaborator Simpolism, who had written:

> [In ancestral environments] these events were potentially cyclical: a tribesman might experience war repeatedly throughout their lives. However, the current state of modern war leaves veterans returning, psychologically prepared for another go at war at any time, but without any real likelihood that they'll be sent back out in the field… the developed priors become useless, rather than necessary preparation for the next conflict. We can also consider how ancient tribes may have handled "bad" prior formation by considering ritual experience.

> The sacred, the psychologically powerful, as a means of restoring a more "normal" psychic equilibrium.

These dynamics are important to establish for coming sections on fashion, anti-inductivity, the contextual nature of heuristics (such as surrogate measures, metrics, and markers), and the fit between environment and heuristic which—given enough time and environmental stability, emerges naturally through evolutionary cycles—but which is forceably uncoupled when adversarial agents seek out environmental manipulations that degenerate the evaluative abilities of the heuristic.

4.3. SURROGATE MARKERS

Darwin Ortiz, *Strong Magic*:

> In the heyday of the big con in this country, professional con men would often meet some wealthy businessman and, within a couple of hours, succeed in convincing him, without any collateral, to turn over large sums of money to them. These well educated and highly intelligent businessmen were wiling to put their trust in complete strangers in large part because the con men were able to convince their victims, purely through their dress, grooming, and demeanor, that they were the same type of men as themselves and therefor trustworthy. These con men know that all of us draw firm conclusions about others based purely on what we see.

In classic game theory, and the formal games which classic game theory analyzed, most variation in the environment is bracketed. A prototypal example is chess, where a carving

nick in a bishop is irrelevant to gameplay, and minor variations in piece position within a given square are "rounded off" or not even noticed. All black bishops (or white queens, etc) are fungible with one another, like paper currency; both game rules and player strategy are unchanged when moving from one chess set to another. The informal games which dominate ecology and sociality, however, feature no such fungibility; any difference may make a difference, and any sensory pattern that correlates with strategically relevant states may be seized upon. S.I. Hayakawa, *Language in Thought and Action*:

> We may infer from the material and cut of a woman's clothes the nature of her wealth or social position; we may infer from the character of the ruins the origin of the fire that destroyed the building; we may infer the nature of the Soviet Union's geopolitical strategy from its actions across the globe; we may infer from the shape of land the path of a prehistoric glacier; we may infer from a halo on an unexposed photographic plate that is has been in the vicinity of radioactive materials.

The relationships between expressive cues (surrogate markers), and the deeper, more meaningful qualities they imply, are known and therefore optimized around. Hotel Concierge, in his essay "How To Be Attractive," writes:

> Consider: "I saw her from across the room, and I immediately fell in love." Fell in love with what? "She had these big thick-rimmed glasses…and an impish smile… and we're holding hands, and it's the fall…" Right, part for the whole. She had big glasses, so you typecast her into the story you've run through your head a thousand times, the story repetition has lodged in your

> unconscious Id. She had big glasses, so she was the type of girl you could love. "No, you don't get it—she looked like the girl of my dreams." Exactly. Of course, she knows this, which is why she chose the glasses.[11]

Sometimes, the public metonyms chosen are more or less honest self-representations. One cannot find cooperative solutions—strategies that are in both players' interests—unless each player has a sense of what the others' interests are. One thus has incentive to selectively reveal, honestly, one's preferences and desires, in order to find compatible allies. This is what makes such metonyms worth interpreting at all—individuals in purely adversarial situations have no incentives to honestly signal, and thus all of their signals ought to be treated with deep suspicion or ignored entirely. The more cooperative the situation, the more reasonable an expectation of honest self-representation is:

> One look at [our coffeeshop Deschanel[12] Doppelgänger] predicts 25% of her personality. She has a Macbook Air, thick glasses, and a floral dress: for some reason, I doubt she's voting [Ted] Cruz 2016. She unconsciously holds her features in line with her default mood: bored, shy, bubbly, bitter. So you walk up to her and say hi and within ten seconds she knows 50% of you, or at least the person you will be around her. First date will put it up to 90%. There are an infinite number of details for the two of you to share, but the name of her childhood stuffed animal has no predictive power, while the way she pauses and inhales before each sentence tells you

11 In other words—Goodhart's Law, but for vibes. Block quote from Hotel Concierge, 2016.

12 Zoe, of *New Girl* and *(500) Days of Summer*.

exactly her insecurities. Put another way: once you have your first fight in a relationship, you know how every other fight will go.[13]

And yet, while in the abstract and long-term, honest self-representation might be the best strategy for finding a stable and high-compatibility alliance, it is incredibly clear that individuals do not self-represent with radical honesty and accuracy in informal selection games. Often individuals wish, even when it is long-term misguided, to present the most pleasing and plausible self-representation they can muster in order to maximize their chances of passing an informal selection game. In online dating, each player would prefer to have accurate information about their prospective date—to know, in advance of a scheduled date, his partner's true interests and appearance. He will feel betrayed if they find these traits have been misrepresented, that he has perhaps wasted his time by selecting the other for a date. At the same time, each player is incentivized to represent themselves as positively as possible, in order to maximize his chance of being selected in the first place. Thus, a balance between short-term and long-term (games of selection, games of sustenance) must be struck which does not stretch reality too far—to both prevent a backlash of surprisal, and preserve plausible deniability in the face of skepticism.

The diagnosis of such behavior as "defecting" or anti-social is complicated by situations in which an evaluated player believes the evaluating party is biased as a selector in a way which hurts the evaluator's own best interests. For instance, a manager hiring a new employee may have a prejudice against a kind of employee which would cause him to avoid hiring

13 Ibid.

an employee of that profile; if an applicant sincerely believes they are the best candidate for the job, but that this fact will be improperly masked by some fact about themselves, then they may benefit both the selecting party and themselves by misrepresenting themselves in order to be hired.[14]

Indeed, because informal games are guided by implication, connotation, symbolism, tone, and vibe—rather than the quantitative surrogates which characterize many formal games—honesty is a more slippery concept; the meaning of symbolic statements is frequently ambiguous, and one cannot be "caught" in a strategic misrepresentation of ones "vibe" the way one can be caught lying about years of professional experience. (At most, these metonyms can be called "misleading.") In other words, both self-deception and general dissimulation profit off the semantic vagueness of the informal game.

4.4. TYPIFICATION

As we have seen with the *Hotel Concierge* passages, typecasting structures agents' strategic reading of one another, as well as their deployment and interpretation of surrogates. Expected and common types organize incoming sensory data from lower to higher levels of abstraction. In a hermeneutic loop, we advances a guess as to type ("whole") through deference to part, and then continually evaluate those parts in light of our type estimate, which is held with varying degrees of provisionality, and which informs the search for "confirmers." These types are guide not just interpretation but action: once

14 Complications like these are part of why the simplistic cooperation-defection paradigm is ill-fitted to the analysis of real human behavior.

an encounter object, agent, or situation has been typified, the typifying agent can begin using their stock of knowledge to reason about the subject and determine strategic fit. In a very gross simplification, a type can be thought of as a set of "rules"—procedures and expectations—for the type by the typifier. This ruleset is highly pragmatic, that is goal-oriented, such that individuals or objects are often classified by whether they hinder or help a given project (whether they are obstacles or assets).

Christina Marinakis, an expert in juror selection, describes various selection strategies through the lens of types:

> I'm not just thinking about who's going to be a good or bad juror for my case, but who's going to be a leader in the deliberation room, who's going to be a follower, who's going to be what we call a consensus builder, someone who's going to try to get everybody to agree. Oftentimes you think teachers, they tend to be consensus builders. They try to get people to negotiate. You're also looking for people who are what we call contrarians. A contrarian is someone who will always challenge the status quo. They like to play devil's advocate… I'm looking at how jurors interact with one another, who's having lunch with who, who's talking with whom in the hallways, who's opening the door for everybody, passing out pens, that person's probably going to be someone who's a consensus builder. Or people who are making jokes who other people are laughing, that person has a possibility of being a leader, who respects whom?[15]

15 Elwood and Marinakis 2018.

4.5. FETISHIZING MEANS

How often we come to think of some means as sinful in its own right, solely because—in some contexts, common to us—it leads to undesired ends. How often we come to think of an ends as undesirable in its own right, solely because—in some contexts, common to us—it stems from sinful means. How often we come to think of those tastes or affinities as dysfunctional, solely because—in some contexts, common to us—they signal underlying dysfunction. How often are table manners converted from statistical signal to autotelic value. All is reified, all is de-instrumentalized, all is taken from cue and clue to thing in itself.

These dynamics of qualitative surrogation—an informal association between some metonym and some deeper quality—can be seen in many similar confusions sometimes classed as "fetish." To fetishize, here, signifies the treatment of a provisional means or marker as an end *in itself*—or, relatedly, the taking of some incidental feature of an instance as an essential feature of a class.

Such dynamics have also been called "cargocult." An origin myth follows: During the Second World War, American troops airdropped massive amounts of food, weaponry, and supplies onto the Melanesian islands as part of their island-hopping campaign in the Pacific. To islanders, long isolated from industrialization, the wealth and abundance of these drops were interpreted within a mystical, quasi-religious framework. When the war ended, and the airlifts dwindled to a stop, cults emerged among islanders attempting to ritualistically summon more supplies. Lacking an understanding of the core mechanisms behind the airdrops—a world war, mechanized flight, the Allied island-hopping

offensive—these so-called cargo cults began constructing imitation runways, dressing like U.S. soldiers, and praying that supplies would come without success.

In a cargocult, classically, common concretia become associated with the abstract class they sometimes instantiate, such that they are taken as necessary and sufficient identifiers of the abstract class.[16] Appearances are mimicked in the hope that function will follow—a tactic which is successful proportional with the degree to which a domain is extrinsic and therefore optikratic; in popular parlance, *fake it til you make it*. It is an emphasis on skin over bones, surface over structure. It wears the clothing or trappings of the thing mimicked, but does not have its musculature; it is a p-zombie; it is the opposite of mechanistic thinking.

Selection and discretionary principles often become fetishized when they undergo a transformation from consequentialist instrument, or useful heuristic, to deontological imperative. A game-like approach by a player can appear like, but not in reality embody, a fetishistic mindset insofar as it takes such principles as provisional deontologies—that is, acknowledging that the social body has come to fetishize certain styles of behavior or sets of action-reaction (stimulus-response) patterns as inherently ethical or inherently preferable. The gameplayer recognizes how important kinds

16 Objects or experiences which are strongly associated with, or especially conspicuous examples of, an abstract category will often find themselves "essentialized"—not in the sense that their detailed uniquenesses are reduced or overwritten by their categorical identity, but in the sense that these categories are reduced to their exemplars' uniquenesses. A speaker who has such a relation to language behaves more similarly to a neural network than to a correspondence theory of truth: the structures and associations of words, and structures to speakers, and structures to roles, guide the activity almost completely.

of reputation and credibility are, and the consequences of breaking fetishized principles, and acts accordingly.

It is inevitable that, once a means is committed to *as* means, it begins to functionally resemble an end. But the degree to which this end is held lightly, and its necessity re-evaluated in the light of changing circumstances, matters quite a bit. As the situation drifts, such that previous heuristics cease to apply, poor players will continue applying the dated heuristic even as it has poor fit (pragmatically-functionally) with the new set of circumstances. We can call this tendency *heuristic stickiness*, or *sticky heuristics*, after the well-known stickiness of market prices.

Sam Fussell, son of Paul Fussell, left behind his literary and law aspirations for body-building, out of *fear*—a fear of life in New York City in the 1980s, a fear that were he to be mugged or assaulted, he would be helpless to resist or stop it. He begins lifting to lose his fear; at some point, the singular logic of building takes over, becomes his fetishized, autotelic ends, rather than his means. Instead of the hard problem of solving the holistic "bravery" issue, he gets lost in a single, value-clear reward function, with quickly diminishing returns to the larger bravery goal. This fetish he casts as *life-denying*:

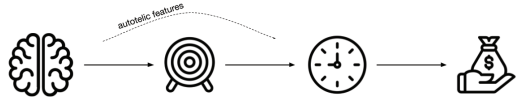
> My lifting was life-denying rather than life-affirming. It didn't have to be lifting or muscles, of course. It could have been tax law or eighteenth-century English literature or arbitrage—anything where the obsession precluded all else. I was as twisted, warped and stilted as a bonsai tree. Another of life's miniatures.[17]

17 Fetish is often accompanied by what C. Thi Nguyen calls *value clarity*, where a simplified goal framework is seductive and emotionally palliative in

The visible payouts of today shape the elected production of tomorrow

The psychosocial incentive matrices that are *attached* to an activity *structure* that activity, by influencing its goals.

Traditional model



IDEA

A possibility bubbles up, perhaps subconsciously motivated.

If deemed intuitively **interesting**, it may be kept or considered.

TELOS

Query: What is are its possible functions? What role does it play?

Audiences are **simulated** for possible response; previous encounters with works are recalled and considered against the telos at hand.

SPECIFICATION

Query: Is this telos worthy of selection ("follow-through")? Is it interesting?

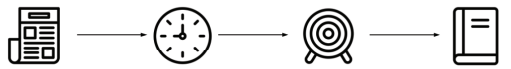
Considerations are primarily situational, relating to the **fit** among parts, between part and audience, etc.

PAYOFF

The fantasized selfish and selfless rewards which are chiasmically **intertwined**. The primary reward is recognition.

If we ask one who loves his job, do you do it for wages or work, what could he say but "both"?

Bourdieu's model



PAYOFF

The public **recognition** of specific literary accomplishments are observed and internalized in the (psychological) incentive matrix of the would-be author, e.g. formal innovation, storytelling, ethics.

SPECIFICATION

Possible paths are considered on the basis of **fit** to: history of the field, current sociopolitical mood, personal aptitude, peer effects, etc. Ideas are pursued if the haver believes they will be rewarded.

TECHNE

How might each function be implemented side-by-side, and knit together?

The legistical process of assembly.

WORK

The finished text, which has been authored by and through the many pressures of its process, which also provide the pretense to write.

Pictured: the role of symbolic capital payouts in shaping the production function of the literary field.

We watch others' strategies to learn from their success or lack thereof; we imagine others' responses to our performance.

Surrogates or expressive technologies which are publicly seen to be rewarded exert an attraction field, while those which are seen as punished exert a repulsion field.

Fetish both of means and metonyms often occurs because it is easier to identify and evaluate the fetish object than it is to assess the original ends. In this way, fetish is a classic surrogation problem. It is *simpler* to track one's objective, numeric weight, bicep size, and bench press than it is to track something as vague, subjective, and complex as "safety." Thus, we are seduced into optimizing the more game-like quality—seduced away from the original ends, and thus begin to ignore the more holistic work which something like "a feeling of safety" might require—for instance, a complex mixture of psychological work, lifestyle adjustments, self-defense training. One frequently sees disagreements that began, ostensibly, over the "territory itself"—the nature of gender, or diplomacy, or immigration—devolve into semantic arguments over the definitions of terms, in part because such disagreements are simpler to resolve and thus more tractable. But by departing from our original goal, and into a side show, we distract ourselves, never to reach the destination we originally desired.

To the consequentialist, deontology itself is a form of fetish. What is morally relevant, always, is human outcomes. As in rule-based utilitarianism, rules may be provisionally implemented on the basis of best accomplishing human flourishing. But the rule is always purely instrumental; it has no value other than in the outcomes it achieves. Deontologies, because they bear a relationship to morality that is roughly analogous to the relationship of letter to spirit, inevitably fall prey to many of the same problems. No elegant compression

its simplicity. Constructed game worlds such as chess or *Super Smash Bros* have significantly more value clarity than real-world games such as running a company. Value-clear games quell our anxiety around which ends ought to be worked towards, and reduces the cognitive space of game-play to pure implementation (*how* instead of *what*).

can adequately and properly address all possible contexts in which it might be applied—this is definitional in the concept of "compression," hence Garfinkel's et cetera clause, or the failures of conceptual analysis in analytic philosophy, which believed it could generate concise, robust definitions that could pick out all legitimate instances of a concept (e.g. knowledge) while excluding all non-instances.[18] A deontology may provide generally good guidelines, but fail as proper moral guides in critical edge cases. In such cases it seems best to abandon rather than adhere to the strict rules; this, of course, introduces the problem of discretion that in institutional setting is referred to as "bureaucratic discretion": if individuals are empowered to discard or rules as they see fit, then it becomes difficult to litigate what are "appropriate" discardings and what are inappropriate.

Alternatively, this conversion of instrument to ends can be conceptualized as a myopia, a nearsightedness. This myopia is what makes games without serious stakes (e.g. casual parlor, video, or board games) work, since it allows us to momentarily forget we are playing just for fun. It is also what causes us to get carried away by a competitiveness which damages our relationships, or injures the good feeling of a social gathering over a trivial game outcome, over symbolic points that do not significantly cash out socially. (Certainly, winning such a game in a cheap way, and causing social friction, is more costly than it is lucrative.) Put simply, we are prone to forgetting the big picture, embedded in tasks and context windows as we are.

18 In other words, the family resemblance frame we are using to define "surrogation" is premised on the same fundamental representational dynamics that surrogation as a phenomenon is premised on.

> It's also about not making scoring your obsession. Otherwise, you're Gollum and the record is your Precious. The real goal is to win games so that you win championships because you want to please the fans who pay your salary and cheer you on game after game. Fans would rather see you win a championship than set a scoring record.
>
> Kareem Abdul-Jabbar on the NBA scoring record

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 and distributed on cassettes well into the mp3 era. The logic for this association was relatively straightforward, if not premised on costly signals exactly, 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 performance 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" (indeed, arguably the opposite). But 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, are 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. The concrete is a surrogate for the abstract pattern which birthed it. As environment drifts, the stranded concrete no longer bears a serious claim to representing its mother.

It is 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 to the exaggeration of public outcry—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 among beginner artists. In Arthur Danto's book-length profile *Andy Warhol*, we encounter the artist'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…

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). Here that very by-product is lifted out of its context and treated as a goal in its own right.

Amidst these performances—which are often enough to fool consumers and 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 tools. 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.[19]

19 On a case-by-case basis, it is of course difficult if not impossible to determine what, exactly, artists are optimizing for. (Artists themselves frequently maintain a zone of cultivated ignorance concerning their own internal motivations.) A persuasive treatment of Warhol's paint drips, or Riot grrrl aesthetics, would require a book in its own right; I use these examples solely to illustrate a more general failure mode.

One related behavior, which I've informally referred to as "map gaming" in previous writing, occurs when writers manipulate (draw connections or oppositions between) surrogate-symbols in a way that is untethered or ungrounded by connections between the surrogated-symbolized. For instance, co-incidences in our collective structure of representation ("map") may be treated and theorized as if it were identical to theorizing the represented territory itself. So-called proof-by-etymology is a common example, although legitimate cases (i.e., cases where map co-incidence does, in fact, accurately signal territory co-incidence) abound.

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's future. We can call its effect *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 negative indicator, the work as zombie art animated by the hungover associations of eras past. Literary critic 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).[20]

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?"[21] We can

20 Jameson, "Experimental Fiction as Genre and as Principle."

21 James is distinguishing between what he calls capital-A Avant and lowercase-a avant work, or Experimental Fiction vs. experimental fiction. The former is essentially genre work: a small subset of possible experiments, associated with e.g. modernism or LangPo, are canonized and enshrined as tropes

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

---

which successive generations of writers feel they must deploy in order to be considered experimental.

commercial. Glam, as Reynolds shows, took the strongest symbols of 60s "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 place selection pressure on—can incentivize or fetishize—a marker or metonym, and agents will remain unaffected by this pressure or incentive. 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?

Scott Alexander, book review of Paul Fussell's *Class*:

> This fits the fables of Early Silicon Valley, where you could wear a hoodie to work because people only cared about how bright you were and not about how you conformed to cultural norms. But (the fables continue) at some point this ossified into a thing where you had to attend interviews in exactly the right kind of hoodie and comfortable jeans, or else they'd identify you as "not a culture fit" and "out of touch with Silicon Valley norms" and deny you a job.

The meta-level is devoured and replaced by its object-level stand-in. The early game's spirit is lost; the late game wears the early game's clothes, but works only by stupid recognition—all its powers of perception, stripped.

4.7. ASSOCIATIVE TAINTING

In the simplest sense, associative tainting refers to the associative, psychological pattern by which a certain symbol becomes associated with one despised (or low-status, or in

a selection game context, merely "unwanted") individual or group, then that symbol itself can become taboo. There is some degree of reification in play, whereby the symbol itself comes to feel "dirty" or degraded, but this is best understood as an abstraction and simplification over the more complex interrelational and historical realities of the symbol, which is pragmatically sufficient to guide individuals to actively dis-associate from, or avoid, the tainted symbol.

Several different realities or game states may be compressed into, or commonly represented by, the same surrogate marker *A*. If it is costly to be mistaken as even one of those game states (*A'*), then surrogate *A* will be occluded from optikratic displays entirely. (Such misreads are often costly because they cannot be corrected; the impressed selector has already dismissed the impressing individual from the selection game, allowing no more evidence to be admitted.) Those wishing to signal alternative game states upstream of compression *A* must search for alternative markers, perhaps one which explicitly denies (by being statistically or causally incompatible with) the possibility of the disavowed state.

4.8. OPTIKRATICS

> ...reality dissolves into appearances and becomes chimerical. Notions of substance get lost in a welter of shadowy images, of staged events, of carefully arranged fronts.
>
> Robert Jackall, *Moral Mazes*

> The first duty in life is to assume a pose. What the second is, no one has yet discovered.

When substance is known through symbol, naturalness is surrendered in favor of seeming natural, authenticity means "a well-integrated social performance,"[22] and we become more interested in "appearing real than being real."[23]

Some have diagnosed this form of modern image-obsession as narcissism.[24] The network of associations and connotations attached to objects, involvements, institutions, and interpersonal relationships becomes the organizing logic of a life lived—prestige sought out through connection with prestige.

I want to diffuse the implied pathology of such a diagnosis. Optimizing for image is strategically rational, not pathological, when others' decisions and behaviors are predicated on their judgments, and those judgments on appearances.[25]

22 Crystal Cultures, Twitter.

23 Eric Hu, interview for *SSense.*

24 See the blogs *The Last Psychiatrist* and *Hotel Concierge*, as well as the podcast *Red Scare*. Narcissism is perhaps best described as a social strategy optimized for shallow, short-term relationships—maximizing the opticratic, at the cost of the intrinsic (and at the cost of creating false expectations).

25 An analogy to game-theoretic defection may be useful. Sarah Constantin writes of Kegan's *Everyone Culture*, and Jackall's *Moral Mazes*: "The basic problem that both books observe in corporate life is that everybody in a modern office is trying to conceal their failures and present a misleadingly positive impression of themselves to their employers and coworkers." As a result of this self-interested impression management, the organization's overall efficiency and output suffers:

1. Employee mistakes become more costly the longer they are covered up.
2. Manager decisions worsen the more they are misinformed by their employees.

Of course, the problem is that strategic optimization is not necessarily "healthy" behavior, in the sense of leading to life satisfaction, and many individuals surrogate their actual priorities, preferences, beliefs, and desires in favor of the image. That is, maintaining "two sets of books" is difficult and taxing, requiring a split in personality. There are two ways to resolve the double-books problem: one can either collapse into image, or collapse the image into self.

I also want to complicate the notion that there is, or was, a world before images. This is simply not the case, in my best estimation. This will be doubly shown in a following chapter, "Evolutions," but I will make the first case here, for how ordinary, banal, and inescapable the optikratic nature of the world is, even beyond the banal philosophical sense that the world is merely sense impressions or phenomena.

Certain situations are almost always presented—in both a cognitive and linguistic shorthand—as other than what they actually are. We say, "People get angry when they have been swindled" to explain why a friend is enraged after being sold a lemon for a car. But more accurate would be to say, "People get angry when they *believe* they have been swindled"—which, normally, we would say only if we wanted to emphasize that we disagree with the angered individual's assessment of the situation. In other words, we point out the fact of perception, separate from reality, only when we disagree with the perception in question. Otherwise, if our perceptions align, they are assumed to be interchangeable with reality. However, it is clearly true that what matters in such

3. Employee investment of time and resources into optimizing appearances (playing "inner games") comes at the cost of investing resources into the advancement of organizational interest (its success at "outer games").

situations just *is* belief and not actuality. When a person is swindled and ignorant of this fact, they are perfectly happy. When they are treated fairly but believe otherwise, they are perfectly unhappy.

Similarly, when we talk about a game, we say something like, "In basketball, two points are scored by a team when one of their players gets a ball through the hoop from within the three-point line, without committing a foul. Three points are allocated when such a performance occurs outside the three-point line." And yet this is not how the game objectively functions! We have ignored the role of perception. The symbolic letter of law is only part of our story—there is also a surplus, the game's spirit, which many players obey above and beyond the minimal letter—and there is a "real" game, a real set of rules, including but not limited to those of physics, which lies below the letter. Gameplay at the sub-symbolic level consists in large part of convincing the relevant (selecting) authorities, tasked to allocate extrinsic rewards such as the (intermediary, state-tracking) "point" or the (culminating) reward: game victory—which is itself important largely insofar as it contributes to the larger tournament situation of win-percentage advancement culminating in the allocation of a championship trophy. And this convincing of relevant authorities is best modeled as a selection game. Thus, one scores points in basketball when one convinces the referees and scorekeepers to allocate points. The institutionally nested referees and scorekeepers who evaluate such appearances are generally bound in allocating points for plausible-appearing reasons, auditable to the officials who hire them,[26]

26 Alexey Guzey, in "Reviving Patronage and Revolutionary Industrial Research," writes: "Grantmakers' planning horizons (note that I'm talking about specific individuals who make specific decisions, not abstract institutions

and to the increasing hierarchical levels of league and then the public which the league ultimately answers to, whose satisfaction is critical to the league's bottom line. In other words, such an officiating setup is characterized by optikracy at both the first- and second-orders: judgments are based not merely on the appearance of the judged subject, but on the appearance of the judgment itself to whichever overseer determines the fate of the judges.

Thus we might say more accurately: "In basketball, two points are allocated to a team when a relevant authority believes that one of its players has successfully gotten a ball through the hoop from what appears to be within the three-point line, without appearing to commit a foul; three points are allocated..." etc. We get away with the shorthand version, which ellides perception and belief, in part because there is no yet discovered (or more precisely, known) move which allows a player to reliably make it appear that the ball has entered the hoop when it hasn't.[27] But it is more at issue with outcome-relevant moves which *are* known to be "fakeable," as is the case with fouls. That is, while almost all social life is optikratic, we only think of games or practices

---

that theoretically care about the long-term) are severely limited by their own career planning horizons and by their understanding of what it takes to work on fundamental problems with little short-term payoff."

27 It is important to point out that the basketball hoop (and the concept of a point, and the rules of basketball broadly) has been intentionally designed so as to be publicly legible, uncontroversial, and "digital" or discrete. (A ball either passes or does not pass through a hoop; there is little in-between.) In other words, the game environment and rule structure have been designed and modified over time to continually facilitate "objective" adjudication. This is a near-universal feature of human coordination, in which the built environment is fixed with clear, discrete breaks (rather than smooth, undifferentiated space) to facilitate clear focal points and decision rules—positions of prominences that minimize ambiguity and facilitate mind reading and synchronization.

as optikratic when the decoupling of perception and reality becomes pragmatically relevant or acutely felt—when degenerate play, in the sense of false appearances, proliferates.

Many individuals, in everyday life (say, in a career setting), claim to do X behavior not for its optics, but because it is the correct and honest course of action. I find this suggestion dubious. I readily concede that many individuals simply do X as a strategy for being seen to do X, and this works to a point; it is a reasonable strategy of play, lessens the cognitive overhead required of double ledger-keeping, and is lower risk than deception. (Others, of course, are more successful putting their energies into pretending.) But let us imagine that the employees of an institution who "actually" perform X suddenly are no longer perceived, by their selecting superiors, to be performing X—indeed, they are now at a risk of being fired (selected out) for performing X, although it is ostensibly their job description and ostensibly advances institutional interests. Most employees at this stage would adapt to behavior which would successfully appear to perform X even at cost of (in actuality) performing X. This transition would be largely unconscious, as conscious knowledge of dissimulation is a liability within the institutional game in addition to causing psychic dissonance. Some individuals might not adapt to this new selection regime, but it is a rare employee who cares more about his employer than himself, and such individuals would be quickly let go (as slackers to boot). Soon, the functioning of the employee pool broadly would have switched to merely pretending.[28]

28 The supremacy of optics can be illustrated by the avoidance, in many settings, of moves which, while perfectly legitimate (spirit-abiding) in reality, might give observers the wrong impression, in favor of moves which are spirit-violating but less apparently suspicious. We can take, for instance, a high school teacher whose daughter is enrolled in his biology class. She may, by all

The only major exception to this rule which I am aware—cases where the majority of employees would continue performing X out of integrity even at personal cost—are those in which the external goal of the wrapping institution are also the highest priority goals of the employees, above and beyond the incentives of employment—for instance, volunteer workers in a disaster zone. And at this point, they are no longer arguably in the thrall of the wrapping institutional game or incentive structure.

Broadly, we can say that if a strategic move would be equally effective, were it undertaken as a perfect sensory illusion, then we can call it an essentially optikratic move. It is about appearances, and not realities.

A general might consider his army to be physically blocking off an entrance to the city, but insofar as the enemy, seeing this blockade, chooses to attack at a different point (or to retreat, or bide its time, etc), the blockade has not exerted any real physical reality so much as it has, through appearances, led the enemy to make a different decision. A good system of

---

reasonable intersubjective standards, be far and away the most capable and knowledgeable student in the class; she may work harder than any other student; she may in every way earn the distinction of top student, in test results, comportment, and overall grade. But when the semester finishes, and awards are given out for each class, on subjective bases, to the teacher-determined top pupil, she cannot be given the award. That is, even if no impropriety has occurred and the award is fully merited, it casts too much suspicion on the fairness of the award, and implies too high a possibility of nepotism, to be granted. Some other, lesser student, will have to be selected instead. And if the award is granted automatically, on the basis of grade percentage in the class, there will be quite a problem for the teacher or administrators, since he or she will have to choose between actually violating the rules of the contest, and appearing to violate its spirit.

projectors, mass hallucination by the enemy, etc would have been equally effective.

This points us to the voluntary, unforced nature of most gameplay. Much of our maneuvering is speculative and anticipatory—it is only when things come to a head that (in some games) real force becomes involved—actual strength and skill instead of their appearance. In the meantime, such games are ruled by information, which operates by a very different logic.[29]

4.9. OPTIKSMIZATION AS CARGOCULT

> The important thing, it appears, is that the numbers have the right form.
>
> Tal Yarkoni, "The Generalizability Crisis"

Optiksmization (n): the optimization of appearances.

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

29 See Bateson, "Form, Substance and Difference" in *Steps To An Ecology of Mind.*

contribute to human knowledge. What's more, since we consciously or uncaringly or by necessity surrogate appearance for reality in decision-making and evaluation, these performances frequently do succeed in "flying," perpetuating the optikratic incentive structure.

These dynamics play out in formal institutional games, and on quantitative metrics, as well as with informal, qualitative ones. 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 (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."

5. More Game Dynamics

5.1. SPIRIT, SYMBOL, REALITY

> The problem [of] real life is... moving one's knight to QB3 may always be replied to with a lob across the net.
>
> Alasdair Macintyre

The selection games discussed in earlier sections are optikratic insofar as the decisions of selector-judges are based not on the real fact of merit—the private, expensive- or impossible-to-assess reality—but on the appearance of merit, on optics.

Furthermore, these games are deeply symbolic. By "symbolic," I mean that there is a strict ritual for proper attainment of game goals, which is more narrow and specific than the space of possible solutions. Players are socially conditioned and incentivized to cooperatively stay within a narrow script-space of winning play.[1] This space is never formally i.e. linguistically demarcated but rather transmitted through example.

1 Lantz & Zimmerman, "Rules, Play and Culture: Towards an Aesthetic of Games":

> The rules of extrinsic games are purely social; they exist in people's minds and are enforced by people. "Once play begins, players are enclosed within the artificial context of a game—its 'magic circle'—and must adhere to the rules in order to participate. If you're playing Candyland, who cares which plastic piece reaches the final space first? The other players do, of course."

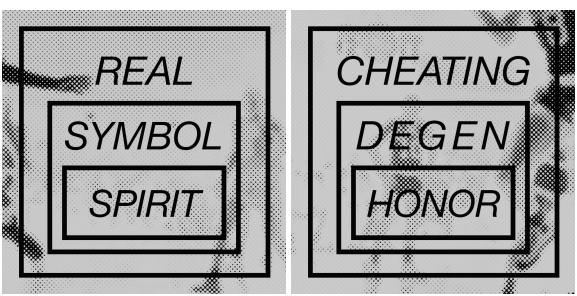
See also Garfinkel's "et cetera principle" for insight into the tacit coordination that underlies game adjudication, from athletics to contract law.

Often there is some disagreement over whether certain tactics are within or outside this space of symbolically and spiritually appropriate play. These disputes reside in the heart of legal praxis and theory; the court case, properly conceptualized, is always a double hermeneutic, an interpretation of both legal text and a historic happening at which the defendant sits center.

These tensions are based broadly in two facts.

The first fact is that, for those competitors who excel at a certain tactic, it is advantageous for the tactic to be considered sanctioned play, while for those competitors who excel at already-sanctioned tactics, it is advantageous to withhold the sanctioning of new tactics which threaten their position.

The second fact is that different play styles (or scripts, or rituals) can lead to different byproducts outside the game, and that both the party hosting the game, and spectators who observe the game, will have prefer certain of these byproducts over others. These byproducts can range from entertainment value to scientific progress, corporate value, the production of new technologies, etc.



We can lay out the levels of a game, and by extension, the kinds of play, according to which level of game they operate on—that is, which level they show allegiance to.

If city law states that five dollars are to be awarded for each cobra killed or captured within city limits, importing dead cobras from outside the city can be considered cheating; importing (or breeding) and then killing cobras inside the city can be considered degenerate; and actually hunting local cobras can be considered honorable. Judgments as to whether a move is degenerate or law-breaking are either moral judgments or morally neutral, depending on whether one takes a perspective from within or without the rule system; such judgments are extensions of and defined by the normative frame coordinated by the game, its legal doctrine, and its social norms.[2] Degeneracy can be more damaging to a game's spirit than outright cheating;[3] cheating can be performed

2 Venkatesh Rao, *The Gervais Principle*: "Effective Sociopaths stick with steadfast discipline to the letter of the law, internal and external, because the stupidest way to trip yourself up is in the realm of rules where the Clueless and Losers get to be judges and jury members. What they violate is its spirit, by taking advantage of its ambiguities. Whether this makes them evil or good depends on the situation."

3 So-called degenerate play is, however, more complex than its label suggests, in that it only rarely brings an end to play, and more typically evolves gameplay in a direction which emphasizes different kinds of skills and abilities. Salen et al., *Rules of Play*:

> When it was discovered that Pac-Man could be played by memorizing patterns of movement instead of through improvisational moment-to-moment tactics, player reaction fell into two camps. Some frowned on using memorized play patterns as a violation of the spirit of the game. Other players, however, capitalized on patterns in order to get higher scores. These pattern players did not consider themselves to be unsportsmanlike at all: they saw themselves as dedicated players who had simply found a better (and more demanding) way to play the game.

out of a desire to preserve the game's spirit (by violating its letter).[4]

If al-Ḥarīrī, quitting the countryside, now enters a chess tournament seeking the prize money, he has multiple avenues for obtaining it: win the necessary matches—which is to say, persuade the tournament organizers of his merit so that they will voluntarily hand over the reward—or to take it by force (e.g. killing or incapacitating the organizers and physically seizing the money)—or those moves which are somewhere in-between physical seizure and voluntary transferral, such as bribes and threats. And of course, the avenue of winning matches does not preclude cheating—more difficult in chess than cards, but always possible.

To give an example from selection games, we can consider college admissions. Columnist Michael Wolff, in his 2006 *New York Times* review of Daniel Goldin's *The Price of Admission*, synopsizes:

> Golden tells us that the admissions process, at least at the 100 top colleges and universities, is not a meritocracy—and exactly who thought it was?—but a marketplace. Every spot is up for bid. Some people bid with intelligence, which has obvious worth to the institution; some with cold cash, with its certain value; and others with the currency of connections and influence and relationships that serve the institution's interests.

4 This type of law-breaking is relatively rare because the potential cost to the individual—social sanction, expulsion from the game, imprisonment or death—result in few individuals willing to undergo such risk for pro-social (but ultimately selfishly unproductive) outcomes.

None of this is to imply that individuals make purely rational decisions of self-interest on a case-by-case basis. Insofar as we are mesa optimizers, optimizing within a changing world, many of our inclinations will have poor fit with a given game environment. We carry with us genetic inclinations—for instance, we may not be totally comfortable in the evolutionarily novel state of anonymity, and may protect our reputation even while around strangers—as well as cultural conditioning, a conscience, force of habit, etc. We employ surrogates, basing our own behavior on the behavior of others, and avoiding anti-social behavior out of an outsized fear of the consequences of being caught. In the face of a novel problem, we will often pick the first approach that comes to mind and stick with it until it ceases to satisfice. There are many complications and shortcomings of a rational actor model, which cannot be rehashed here. We need only believe that rewards function as attractors; that "solutions" to incentive structures often spread through mimesis; and that only a small portion of an overall population need defect in order to degenerate a game past playability.

Here, the symbolic game of admissions, in which applicants are considered on the basis of their academic achievements, is revealed as a nested, public-facing front of a larger real game. The amount parents of applicants are willing to pay—typically in the realm of tens of millions USD—testify to the stakes of selection.

In most chess tournaments, however, the rewards (financial and otherwise) are not worth the risk of a prison sentence—or even the social humiliation of cheating. Most players are members of a chess community, where reputation is crucial to long-term belonging, acceptance, and social status. (In other words, the situation is quite opposite to that of undergraduate admissions.) Therefore most players end up playing according to the symbolic game; furthermore, since professional chess is a game culture in which letter and spirit of play are identical (see §"5.4. Sirlin's Scrub", p. 143) the symbolic and spiritual game are similarly identical; all available moves are considered fair and a cultural philosophy of "total war" entails a pure pragmatics of play.[5]

Still, the real game of intrinsic, mechanical reward has not disappeared—merely, most foreseeable avenues of securing a payoff that diverge from symbolic play have been made difficult and therefore risky or expensive. The host nation has invested considerable resources in a legal system which punishes crimes and catches perpetrators. Formal regulations and informal reputation systems among tournament officials makes the acceptance of bribes, or other corrupt behavior, costly. The game state of the tournament en toto is publicly visible by all participants, audience members, and officials or proctors, all of whom can notice and testify to differentials between symbolic play and outcome. It is a game easily surveilled and easily litigated, with fewer degrees of surrogation than most.

Of course, were the financial rewards of a tournament high enough, or the blocking (or "counter-") moves of the various institutions and social bodies (governments, chess

5 *cf.* "All's fair in love and war."

associations, officials) less effective, symbolically void play would be less risky or expensive to accomplish, and robbery or systematic cheating may well rear their head. Some players would likely continue to attempt to win via the symbolic game, others through the real game, depending on their capacities and inclinations. (Even if many tournaments were robbed at gunpoint, Kasparov's best strategy for winning the reward money would remain symbolically observant play.)

Thus, what is considered symbolically "in bounds" or "out of bounds" is of great importance. In WWE (World Wrestling Entertainment) productions, acting and melodrama by the athletes is considered in-bounds, in part because the central aboutness of the game is or has become entertainment and drama.[6] In most professional sports proper, acting (for instance, flopping on a foul) is considered degenerate or cheap.[7] Sometimes it is within the rules, if dishonorable; sometimes it is "against the rules,"[8] but the difficulty of attributing intentionality to e.g. a fall—determining what is the "prop-

6 In the early 20th C, professional wrestling was more of a genuine sporting competition than a scripted performance. It serves as an interesting case study of how a game's spirit can change over time.

7 Noting a similar behavioral incentivization, Natalie Wynn (among others) has argued that contemporary cultural norms subsidize and encourage individuals to self-present as victim: "We've all become Italian football players writhing on the ground in fake agony" (Twitter 2021).

8 "Against the rules" in scare quotes insofar as, while officials may make claim to incorporating intentionality into decisions, we know better: only intent's lossy, ambiguous surrogates are on display. Further, I believe that even intentionality is itself a surrogate for understanding an individual's behavioral algorithm, in order to predict future behavior. Insofar as an action is "accidental" or "unintended," it cannot be expected to be performed by the given individual again at a rate higher than chance. Insofar as an action is deliberate, it reflects an attitude, orientation, or behavioral algorithm upstream of future, similar actions. See also prison terms, displays of contriteness, and the

er" amount of reaction to a push or an elbowing, vs what is dramatized—is nearly impossible, making it de facto part of the game. Foul-drawing—for instance in basketball, shot attempts which are not sincere efforts at scoring, but rather efforts at creating legitimately unlawful contact by a defender—are somewhat more controversial. Many commentators, while pointing out that a shot attempt is insincere, will also argue that a defender should have "anticipated" such a move from the offensive player, and adapted their play style accordingly—in other words, that the move is fair because it has been routinized to the extent that it can be expected, and thereby incorporated by the defender into his strategy. (This being one of many examples as to how common knowledge and the concept of fair play coincide via the metaphor of a leveled playing field—see Hammurabi's Stele.)

This spirit is varyingly arbitrated and constructed by any agents who administer (indirectly influence, or directly determine) the reward function. These agents can consist of the game designers, hosts, audiences, other players (since peer approval is one of many goals players optimize toward) etc. Even when designer intent has no programmed relevance to the real reward function, onlookers will often defer to (their impression of) designer intent, or else use speculation as to intentionality, as the basis of informal spirit arbitration. The spiritual aboutness often emerges from what onlookers or sponsors find valuable in the game already. In the EmpLemon *Super Smash Bros: Melee* documentary *there will Never Ever be another Melee player like Hungrybox*, the narrator argues:

---

oft-observed tendency of intentionality to be hidden even from the intending individual himself.

> For many fans of the the game, Jigglypuff [—the character Hungrybox plays as—] represents the antithesis of everything that makes *Melee* great. She requires lower technical skill than the rest of the high-tier characters. She's floaty and hard to combo. Her playstyle is inherently slow, campy, and defensive. Spectators often accuse her of being boring to watch and play.[9]

In zero-sum games, out-of-bounds play naturally comes at the cost of in-bounds play, which makes rules and rule-following (even on the margins) a central concern of players. Players have limited time, energy, and resources which they must allocate; there are a limited number of—or preferential ranking of—admission slots e.g. in play-off tournaments, limited grant foundation funds, college admission or hiring spots, monogamous partnerships, etc. Very simply, those players who exclude from their optimization equations those criteria which are not strictly necessary to a desired outcome will outcompete those who take on such additional constraints. By extension, players who "specialize"—or narrow their goal, e.g. desiring only financial success while remaining indifferent to the approval of peers—thereby outcompete, at the given goal, those who "try to have it all."

This disadvantaging of in-bound purists provides a psychological rationale which leads even honest players to adopt degenerate or illegal tactics. Major League Baseball has

9 One of Hungrybox's main rivals, Leffen, has been consistent in his public statements that Hungrybox's play style is degenerate, arguing that its defensive, slower pace is "killing" *Melee* as a game by driving gamers away (despite viewership and participation numbers steadily increasing during the period of Hungrybox's dominance). Considered outside the symbolic reality of subcultural space, such statements ought to be considered an attempt at increasing the social tax levied against Hungrybox's play style.

recently been scandalized by its pitchers' coating of baseballs with resin in order to pitch at higher speeds; a minor-league pitcher, when interviewed, relates:

> "The calculus is whoever gets outs better gets to play major league baseball," says the NL reliever who says he uses Pelican. "There's some guys that might have a moral dilemma about it, but I'm not one of those guys."[10]

Playing honorably—that is, in compliance with both the symbolic rules as well as the symbolic spirit—is a form of socially cooperative self-handicapping. It is incentivized primarily through reputation (social sanction) and conscience (acculturation). While cheating, in the strict sense, nearly always must be hidden from view, degenerate play often occurs in the open, since visibility to officiating parties does not change the outcome of play, and all else equal, concealment is costly. The advantages to concealing degenerate play are two-fold: first, the lack of reputational cost; second, that any advantage garnered by the degenerate tactic ("exploit," "bug," "loophole") will be erased if the tactic is widely disseminated across the field of play. (Adoption quickly accelerates past the pioneer stage: popularization of degenerate or unlawful tactics makes such tactics less risky for any given individual, similar to the logic of a mass protest, or of attempts to subvert a preference falsification regime.[11]) It is the most honorable players who are of course hurt by these fads, though eventually, a logic of "everyone's doing it," and an embitterment in the face of repeated defeat, tends to sway hold-outs. Even if individuals are not persuaded, they are

10 Apstein and Prewitt 2021.

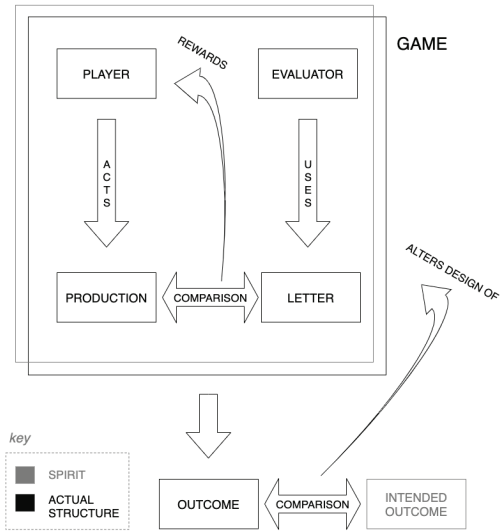
11 *cf.* Timur Kuran's research.

selected-out of the system as less fit. Only the very best players are able to maintain their stature while playing honorably—a costly signal of their ability.

> Says one of the NL relievers: "For us that refuse to use sticky [stuff], we get pushed out, because 'you don't have great spin rate.' Well, no shit, because I don't cheat." [...] At the moment, umpires generally rely on managers to request that they check a pitcher. Managers largely refuse to do so, in part because they know their own pitchers are just as guilty, and in part because they worry their team may someday acquire the pitcher in question. Executives and coaches who personally abhor the practice do not see much benefit in telling their own pitchers to knock it off, knowing that will accomplish little more than losing games and angering their employees. Fringe pitchers tell themselves that everyone is doing it—indeed, that the league's clumsy management of the game all but requires it.[12]

In some cases, public awareness—e.g. that brought about by the excerpted *Sports Illustrated* story—combats such play by bringing awareness to the public, who in turn demand rule changes or else accord less prestige to players known to engage in degenerate play. Often though, publicization merely drives degenerate play underground—e.g. it is taboo in many circles to admit that one's undergraduate admission may have been influenced by a substantial parental donation,

12 Spectators and fans, meanwhile, seem caught up in the thrall of large, round numbers: "We're just doing the same thing we did during the steroid era," says the other team executive. "We were oohing and ahhing at 500-plus-foot home runs. ... A 101-mile-an-hour, 3,000-rpm cutter, isn't that the same thing as a 500-foot home run? It's unnatural."



or that a romantic relationship came about by hacking the algorithm of a dating app.[13]

The argument that a cheap tactic is widely in use serves, first, to normalize the behavior; second, to testify to the greater security of crowds and large numbers (lessening the chance that one is singled out for punishment); third, eases moral concerns by implying that, since, a majority have already defected, individual refusal cannot save the system, and individual defection will not meaningfully degrade it. Rick Singer,

13 See mathematician Chris McKinlay's use of OKCupid, as profiled by Kevin Poulsen in *Wired* (2014).

the college counselor at the center of the 2019 Operation Varsity Blues admissions scandal, used similar rhetorics to convince parents to seek disability exemptions for their (ostensibly neurotypical) children[14]:

> "The Academy kids are getting extra time all the time. Everywhere in the country... What happened is, all the wealthy families that figured out that if I get my kid tested and they get extended time, they can do better on the test. So most of these kids don't even have issues, but they're getting time. The playing field is not fair."

Games can also vary in the extent to which social sanctioning—and by extension, the spirit of play—matters. In some games, a desire for fairness, and the non-discretionary disbursement of reward, lead to all letter-abiding behavior being considered equally valid. This means that systems low in corruption are often high in degeneracy.

5.2. BEYOND SYMBOLS

Chess, and board games broadly, are a valuable foil to real world games, because their manipulated symbols are never asked to stand for anything beyond themselves.[15] Note how untrue this is broadly: The "homicide rate" that a country's citizens care so much about is only so-called, because it is not actually the rate of homicides in the country. On the

14 This abuse of disability accommodations upset several prominent disability rights organizations, who complained that such abuses of the accommodations discredit public legitimacy of accomodations more broadly, and threaten the future ability of disabled students to secure such accomodations.

15 And also because there is no public-private information gap in third-party assessments. (The public position of a piece just is its real position.)

one hand, this fact is trivial; on the other, it is crucial and constantly forgotten. The homicide rate is the rate of deaths which are discovered and then ruled—that is, interpreted as—homicides by law enforcement. This number is obviously correlated with, and influenced by, the actual homicide rate—but it depends also on the number of reported disappearances, the abilities of local police, the cleverness of criminals, etc. We care about the "homicide rate" going down not because we care about police interpretations but because it implies the "real" rate has dropped. (And this is "real" in scare quotes because homocide is an abstraction without natural fact, without natural "joints.") If we learned that criminals had become more adept at fooling police, e.g. with phony suicides, and that this was the cause of the drop, we would greet the news with worry instead of elation. The statistic is merely a surrogate for what we "really care" about, and on which we must rely. Officials whose standing depends on the performance of the surrogate will inevitably end up manipulating it, thereby degrading the strength of the correlation between surrogate and surrogated; this Goodhartian dynamic is the subject of *The Wire*, and its scathing critique of both political optics and stat-padding in public institutions.

Not so with chess. The game state can be easily visually assessed in toto, and the appearance of the board pieces is equivalent with the real state of the game. Because their positions are purely symbolic, there is no schizophrenic reality-impression split which counts. There is some space, perhaps, for misdirection—for implying one directionality in one's tactics, while surreptitiously pursuing another—but this space exists only at the level of futures modeling. There is little role for conceptual or aesthetic interpretation, only

sensory assessment, which is straightforward and rarely a subject of public dispute. There is relevant private information for each player—as in the case of misdirection, the opposing player has an obvious interest in knowing the hidden intentionality of his opponent—but for the purposes of the third-party judge, all necessary information is out in plain view. He will need to make no conceptual inferences about things hidden, about events geographically distant or lost to time. At any given moment, the game state is fully available; the representations of game state are the game state literal; and the "meaning" of pieces and positions is discrete and well-mapped so that there can be little interpretive doubt. Symbol and substance are the same—and this makes all the difference in assessing game outcomes.

Compare an internal affairs team which is attempting to assess the scene of a shoot-out between police and a street gang. There is a selection game between the IA team, which seeks roughly to determine the relevant truth of the shoot-out, and the police officers, who wish to escape being selected for punishment. Even objective questions the IA team may wish to answer—that is, questions with an ostensible fact of the matter, such as who fired the first shot—are lost to the past. There is some ability for on-scene evidence to testify to these questions, but such evidence can have been tampered with, to testify in a way beneficial to the tampering party. And these objective questions are typically themselves surrogates for getting at more difficult, aggregate, and subjective questions, such as whether the officers' use of force was "justified" (spirit, letter). Inevitably, to de-vagueify the concept of justification, certain markers and determinants are formalized, if only through the concept of precedent—but this de-vagueification also makes the judgment more gameable.

In the example of chess misdirection above, we can conceptualize a selection game where each player is selecting from a set of possible moves which they hope will advance their in-game prospects. Each player's opponent has a direct stake in how the player chooses. Unlike in actual warfare, one cannot erect rubber tanks in one region to feign an imminent attack there, while secretly moving one's actual weaponry to different area of the map.[16] A player cannot make it look like he has moved a pawn to E5 when he has in actuality moved it to E4; he cannot really make it look like there are no open rows by which to check his King when there are in fact open rows. But he may be able to leverage his reputation as an aggressive player to upset opponent expectations, or to imply one larger plan of attack while in reality setting up another. That is, both players play with an eye to the future—what they believe past moves indicate about the likelihood of moves to follow—and there is a gap between this apparent future and the actual planned-for future.

What is important here is that, unlike in the case of al-Ḥarīrī and his lion, or of a job interview, the players are not selecting on another (to be eaten, to be hired). Instead—and this, arguably, is the relevant superset of selection games this text deals with—one party has a decision to make (a choice among available options) and another party has a stake in which option is picked; there is some non-trivial ramification, in the selection, for his own interests and opportunities. This, more or less, is the state of the market (choosing a shampoo brand is a selection game, insofar as agents create a product

16 See Allied tactics leading up to the Normandy invasion as discussed by Goffman, *Strategic Interaction*.

which is selected among options by a purchasing agent). It is, more or less, the "garden of forking paths" made famous by Borges. And it is the state of the ecological huddle in which actions have a mutual relevance. Organisms are ecologically connected if their actions affect one another, and modernity has been tremendously successful at extending our nervous systems and interests such that our individual ecologies are vast and global, the buffer between men thinned.

When one purchases a used car, one might arguably say that one has "selected" the car owner; when one purchases a shampoo, one might arguably say that one has "selected" the shampoo manufacturer; that in either case, the situation is not too different from al-Ḥarīrī and the lion. But when a chess player selects his move, or the American military selects a strategy for the Middle East, they are not choosing interested individuals, they are choosing in a way that is *of interest* to individuals. They are choosing based on their impression of the game state and affordances, based on predicted payoffs of one choice versus another, and the interested individuals have clear incentives to influence the choosing party's assessment of game state and payoffs in order to alter this choice.

5.4. SIRLIN'S SCRUB

David Sirlin, former designer and top international player of the *Street Fighter* games, defines a scrub as "a player who is handicapped by self-imposed rules *that the game knows nothing about*" (emphasis added). These rules are an "intricate," "fictious" construct, an idealized and vague set of so-called "principles" defended by notions of "honor" and "cheapness." In *Street Fighter*,

> Performing a throw on someone is often called cheap. A throw is a special kind of move that grabs an opponent and damages him, even when the opponent is defending against all other kinds of attacks. The entire purpose of the throw is to be able to damage an opponent who sits and blocks and doesn't attack. As far as the game is concerned, throwing is an integral part of the design—it's meant to be there—yet the scrub has constructed his own set of principles in his mind that state he should be totally impervious to all attacks while blocking.
>
> You will not see a classic scrub throw his opponent five times in a row. But why not? What if doing so is strategically the sequence of moves that optimizes his chances of winning? Here we've encountered our first clash: the scrub is only willing to play to win within his own made-up mental set of rules.

Sirlin is certainly right, as he goes on to argue, that these notions of honor and cheapness are often strategically motivated. Players may lack an effective counter to a tactic, or be weak at using the tactic themselves; socially tabooing such a tactic gives them an in-game advantage, while upholding their dignity and social standing outside the game. (Social distinction being one of if not the primary aim of all competitive play.) It is rare that players whose success is predicated on a certain tactic will cede that the tactic is cheap or degenerate, since such a concession threatens the social legitimacy and long-term legality of the tactic.

But we see in Sirlin's thinking, as we see often elsewhere, an inconsistent and somewhat ad hoc marshaling of justification—on the one hand, the throw is "meant to be there," in

other words, Sirlin appeals to designer intent to legitimate a given tactic. On the other hand, if the game mechanics allow a given tactic, designer intent is irrelevant:

> If an expert does anything they can to win, then do they exploit bugs in the game? The answer is a huge yes—for most bugs. If you think "no" is a reasonable answer, then you haven't thought this through yet. There is a large class of bugs in video games that players don't even view as bugs; they aren't even aware that they are bugs.
>
> [...] How [Custom Combos] were intended to be doesn't really matter: in the game we have available, they work how they work, and taking advantage of that is necessary to win.

The question raised: well, which is it? And if the logic of being in- or out-of-bounds does not come coherently out of consistent philosophical principles, then is a strategic justificationism, where means (principles marshaled) serve an ends (interpretation of spirit) hopelessly biased by personal investment?

Sirlin is also astute in his observation that there is a common confusion of effort with efficacy, or difficulty with deservingness[17]—a Protestant just-world view in which high-effort play going unrewarded is "unjust":

> The scrub... talks a great deal about "skill" and how he has skill whereas other players—very much including the ones who beat him flat out—do not have skill. The confusion here is what "skill" actually is. In *Street Fighter*,

17 Expressed as folk proverb: "Play smarter, not harder."

> scrubs often cling to combos as a measure of skill. A combo is a sequence of moves that is unblockable if the first move hits. Combos can be very elaborate and very difficult to pull off...
>
> I once played a scrub who was actually quite good. That is, he knew the rules of the game well, he knew the character matchups well, and he knew what to do in most situations. But his web of mental rules kept him from truly playing to win.[18] He cried cheap as I beat him with "no skill moves" while he performed many difficult dragon punches. He cried cheap when I threw him five times in a row asking, "Is that all you know how to do? Throw?" I gave him the best advice he could ever hear. I told him, "Play to win, not to do 'difficult moves.'"[19] This was a big moment in that scrub's life. He could either ignore his losses and continue living in his mental prison or analyze why he lost, shed his rules, and reach the next level of play.

But I wish to argue that, to call or see a player as a scrub, as Sirlin does, is not to note that he plays by an unreal code, but that he plays by a code *the accuser does not recognize as legitimate*. Sirlin—and virtually all players of games—are a kind of scrub. That is, Sirlin and his cohort of pragmatic, "whatever

18 We all, constantly, self-handicap this way in our everyday lives because we are enmeshed in many simultaneous or interdependent games, which is to say simultaneous or interdependent goals, so that "maintaining office culture" and "feeling proud of our work" and "contributing to society" vie with, say, the salary game. This is the value complexity of everyday life which we leave behind when entering the value-clear singlemindedness of contrived gameworlds.

19 An alternate frame for understanding this conflict is as one between deontology and consequentialism.

J.J. Redick: "When I talk about mentality, look, there's the competition part, there's the physicality; you and Jalen [Green] chased me around... I knew what I was in for. But the extra stuff. When did you make that decision? Like, 'You know what? I'm gonna flop here, I'm gonna run into Pascal Siakam, and jump 7' that way.'"

Marcus Smart: [laughs] Well that play was strategical. They'd just put in the new challenge rule; that's a playoff game, so in my mind I'm thinking, if I can get the ref to call it in my favor, what's Toronto gonna do? They're gonna challenge it. Which means they're gonna use their challenge that they cannot have in the fourth quarter, and we still have ours. And it actually worked that game because it was a big play where, it could've went their way if they had their challenge, that would've won them the game and evened the series out. So it actually worked perfectly."

Co-host: Where do you feel like you learned this?

Marcus Smart: "I don't know... I watched, before my time, NBA players doing it; you watch the overseas players coming over here, they started it; it works for them, so you go, 'OK, I'll take a bit of that, put it in my game, see how it goes. There's no better feeling than when you bait a guy into a trap, get him thrown out of a game, get a foul called on him and he just goes ballistic."

Three Four Two show
Feb. 14, 2022

works" players themselves ostensibly follow an honor code which puts implicit bounds on what kinds of play are acceptable—for instance a ban on aimbots, on running code within the game, on outsourcing, on card-counting, on performance enhancing drugs, etc. There may be symbolic rules in the tournament against any of these behaviors or more,[20] and Sirlin counsels that players ought to obey these. But the physical reality of the real game—the rules of the game environment itself, which Sirlin champions—does not *preclude* such banned moves; rather, it imposes harsh sanctions if the player is caught (by vested authorities) deploying them. That is, properly conceptualized in a pragmatic perspective, symbolic rules are only ever gambled penalties.

Sirlin is part of a culture of play which, outside the controlled environments of tournament play, has tacitly coordinated around a set of self-handicappings, using the reasonable, discretion-minimizing boundary of in-game physics (the space of physical possibility) to define acceptability.[21]

20 Sirlin advises that tournament organizers choose only bans that are "enforceable, discrete, and warranted."

21 In meatspace, Schelling points are crucial to the establishment of letter laws, and the logic of evaluative clarity (simplification of a spectrum into is/isn't) may outweigh other considerations (such as spiritual alignment, or pragmatic advancement of host in its external game). But such prioritizations of evaluative clarity can lead quickly to degeneracy, since the pragmatics of the situation lead to Schelling point violations, beginning a "slippery slope" descent. Official MLB rules dictate that no foreign substances can be applied to a game ball, a clear line in the sand—but a line which is perhaps inappropriately drawn, for the purposes of players. From the *Sports Illustrated* cover story:

> Brand-new major league baseballs are so slick that umpire attendants are tasked with rubbing them before games with special mud from a secret spot along a tributary of the Delaware River. Pitchers also have access to a bag of rosin, made from fir-tree sap, that lies behind the mound. Hitters

Like Schelling points, such boundaries form around prominent or conspicuous features of either the environment or our formal conceptual system (e.g. prioritizing numbers like 3, 5, 12, and base-10, as in the "three strikes" rule).[22] Certainly, litigating based on whether a play is spiritually aligned is more difficult and subjective than merely obeying the programmed laws of a determinate system. That is, in a determinate, programmed system, what is possible and what is considered socially, culturally lawful are equivalent: if one is able to, one can and is well within one's rights. This approach levels the playing field and can be easily arbitrated; selectively policing degenerate plays, meanwhile, requires difficult group consensus-making, coordination, and oversight. In the "open" and subjective world of the real, what is possible and what is lawful are not, obviously, the same—that is, the real and the symbolic games are distinct levels instead of collapsed.

---

generally approve of this level of substance use; a pitcher who cannot grip the baseball is more likely to fire it accidentally at a batter's skull.

But it has slowly, and then quickly, become clear that especially sticky baseballs are also especially hard to hit. For more than a decade, pitchers have coated their arms in Bull Frog spray-on sunscreen, then mixed that with rosin to produce adhesive. They have applied hair gel, then run their fingers through their manes. They have brewed concoctions of pine tar and Manny Mota grip stick.

Today, some MLB teams have gone as far as hiring chemists specifically for the purposes of developing "sticky stuff" compounds for pitching, with significant increases in pitch speed and spinrate.

22 In U.S. Constitutional law, "bright-line tests" are defined so as to minimize interpretive degrees of freedom, and thereby make predictable and regular the law's application. These are contrasted with so-called "balancing tests" which weigh many factors holistically. See also "vagueness doctrine."

What is important in considerations of in- and out-of-bounds, legal and illegal play is, ultimately, the pragmatic purpose of the game itself. What skills do we wish to witness or incentivize? What byproducts do we wish created through play? When Sirlin notes that scrubs play with some sense of non-boring, non-abusive, balanced play in mind, he is really pointing to their philosophies of play—philosophies in service of purposes. Sirlin may be willing to deploy tedious, degenerative tactics in order to score a win, but the scrub is not, and this is primarily a question of culture. A victory in a group of scrubs that is obtained degenerately may not, in fact, be a winning strategy, because "winning" in such a culture of play is social as much or more than it is literal in-game victory. These players are not in a "total war" situation, but rather are actively, tacitly or explicity coordinating via their sense of honor; the aim of this coordination may be to ensure a level playing field (i.e., to include all participants) or to maximize player fun. And when the rewards of gameplay, for *any* player or culture of play, are largely extrinsic and social—the recognition of one's peers, within a culture of achievement—then any victory which is not recognized by one's peers is not a real victory.

In Sirlin's culture of play, it is primarily tactical-mechanical skill which players wish to isolate and test;[23] in other play cultures, the scope of tactics may well include or focus on technical (in the sense of programming and engineering) ability—the training of AI programs, or the design of assistant technology for human players, such that these cultures become duels between programmers more than between "players" as we typically understand the word. To such a

23 Sirlin mentions mind-reading, the search for patterns, and the countering of opponent exploits as the thrills of his personal play culture.

culture, Sirlin, with his self-imposed handicap of manning the controls himself, is a scrub voluntarily playing by the imaginary rules of limited war.

Whenever a player at a "higher" level of pragmatic play—that is, closer to the "real game"—encounters a more "honorable" scrub, the scrub will of course carry a distinct disadvantage. One of history's more famous examples is the approach of Union generals Grant, Sherman, and Sheridan in the closing years of the Civil War, which included a no-holds-barred, total war strategy to victory over the South—in direct contrast to the more aristocratic tendencies of both their opponents and predecessors.

5.5. PLAYING GAMES, LEAVING GAMES

Bhagwat's "Playing Games To Leave Games" (*Ribbonfarm* 2014) touches on a common cadence of interpersonal games—that of longer-term, lower-stakes games punctuated by shorter, intermittent high-stakes games. These high-stakes games are typically either qualifying or assessment games—we can call them "entrance games." They are required to gain admission to further levels of "endurance games," for instance, the institutional game of holding a given position and fulfilling its duties well enough to avoid expulsion or perhaps gain eligibility to further high-stakes rounds. Another common pattern is the qualification for high-stakes "crowning" or "title" games via long-term, lower-stake performance.[24]

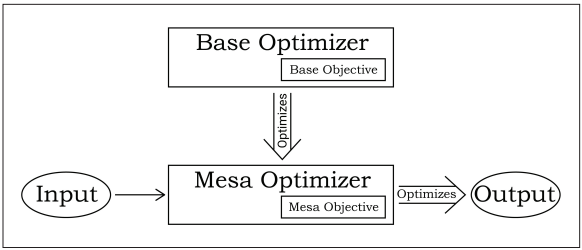
24 Carse, *Finite & Infinite Games*: "What one wins in a finite game is a title. A title is the acknowledgment of others that one has been the winner of a particular game. Titles are public. They are for others to notice. I expect others to address me according to my titles, but I do not address myself with them—unless, of course, I address myself as an other."

Agents often spend large amounts of time training and preparing in advance of high-stakes entrance game—amounts of time orders of magnitude larger than the length of gameplay itself. The lower-stakes training period is often marked by learning, drilling, and the assembly of materials, information, and resources required to succeed in high-stakes competition. There are often distinctions in acceptability between "on-" and "off-court" behavior, and the rhythm of entrance and endurance play is often seasonal, or cyclical.

College admissions, funding and acquisition rounds, hiring processes, sporting competitions, art and musical performances, dating, and warfare all match this rough cadential profile. The lower-stake play can include smaller selection games like alliance-building, resource acquisition, tactical development, team bonding, etc—but these are only partial contributors to high-stake success, and often happen well in advance of the high-stakes situation.

Intriguingly, Bhagwat argues that engineering mindset is antithetical to effective gameplay. "Work importance is relatively constant" in engineering jobs; there is little use for drilling or structured practice; and "engineers are renowned for failure to recognize the important game situations in their midst."

> Patrick McKenzie has made a living out of teaching engineers to play the SaaS pricing game. Steve Blank has done the same for the customer development / sales cycle game. Engineers-turned-startup-CEO are often blind to game dynamics in the corporate budgeting process. A large part of tech's diversity problem is because



engineer interviewers aren't sufficiently attuned to the game dynamics of the hiring process.[25]

One possibility is that the engineering and systems perspective—with its ethos of production—clashes, or is somehow partially exclusive, with strategic skills or mindsets, such as the performative or optikratic nature of strategy, the theory of mind modeling required, and the "surfing of uncertainty" necessary in high-complexity, open-world domains. Where designers, managers, HR departments, and PR agents play extrinsic games with other people—for instance, the anti-inductive game of fashion—engineers tend to play intrinsic games against physical environments, which do not adapt intelligently to the engineers' interventions (i.e., such games are not in fact strategic, because there is no mutual modeling and no anti-inductivity). Some studies have indicated there may be a further link between extraversion and novelty-seeking, on the one hand, and introversion and desire for ritual. The closed world of machines and physics is more predictable, and less anti-inductive, than human psychology. But this is unqualified speculation. The important, generalized dynamic is that "individuals... who have the time and talent to perform a task well may not, because of this, have the

25 "Playing Games To Leave Games."

time or talent to make it apparent that they are performing well."[26] Insofar as the surrogates of a selection game diverge from the qualities they hope or claim to stand for, there is an according divergence in skillset.[27]

5.6. MESA OPTIMIZATION

In their 2019 paper,[28] Hubringer et al introduce the concept of mesa optimization: a "framework that distinguishes what a system is optimized to do (its 'purpose'), from what it optimizes for (its 'goal'), if it optimizes for anything at all."

This frame can help us expand our concept of institutions' internal games from constructed incentive structures to evolved selection mechanisms. Mesa optimizers are selected for by "base optimizers," and "inner alignment" refers to an alignment between the base and mesa optimizer—for instance, natural selection is a base optimizer selecting for

26 Goffman, *The Presentation of Self in Everyday Life*.

27 Jessie Bernard in 1954's "The Theory of Games of Strategy as a Modern Sociology of Conflict" (*American Journal of Sociology*) makes a similar case: "It may even be that the scientist is peculiarlary unfitted for inventing good strategies in a conflict situation. He is not accustomed to dealing with forces that fight back, try to deceive, or deliberately becloud the situation. The story is told that a medium once invited the distinguished Harvard psychologist, Dr. William McDougall, to investigate her performance in order to demonstrate the validity of her supernatural skills. The magician, Houdini, stepped in and volunteered to do the investigating instead, on the grounds that a scientist could not detect the tricks but that he, a fellow-trickster, could. It may be that the scientific habit of thought which the profession of science selects and cultivates operates on a plane where a moral and intellectual and philosophical atmosphere unfits the scientist for the creation of conflict policies."

28 "Risks from Learned Optimization in Advanced Machine Learning Systems."

reproduction; organisms are subject to the base optimizations of natural selection even as they themselves may have goals which only partially align with the base optimizer's goals. Modern non-reproductive sex is an example of a technologically-enabled uncoupling between reward from the perspective of the base optimizer—natural selection—and the perspective of the mesa optimizer—a human being.[29]

The authors stress that not all optimized systems optimize (i.e., are "mesa"). A bottlecap is optimized to selectively contain and release liquids from a bottle, but it is not itself an optimizer. It has been optimized by human beings (much like, say, our food has, be it through recipe improvements or plant and animal breeding). A system is an optimizer only "if it is internally searching through a search space (consisting of possible outputs, policies, plans, strategies, or similar) looking for those elements that score high according to some objective function that is explicitly represented within the system." This is a formalized version of our player within a game, and we will focus on optimizeds-that-are-also-optimizers—on "mesa optimizers"—because we are interesting first and

29 Were we to anthropomorphize evolution—as if there were a designer involved, such as the Christian God—then non-procreative sex could be described as degenerate, since those who engage in it extract from the reward function without accomplishing the actions or ends that the reward function was designed to motivate. The problem, of course, is that evolution uses a surrogate for reproductive attempts—sexual intercourse. As we will see, in one context (pre-technological times) this proxy was so tightly coupled with the proxied behavior that the two were functionally equivalent, and thereby the former could reliably select for ("on behalf of") the latter. Technology has changed the environment, and by extension available player strategies, and thus uncoupled the surrogate from what it stands for.

As an alternative tack, non-reproductive sex can be described as degenerate insofar as it literally degenerates—brings an end to—the infinite game of genetic survival which characterizes natural selection and our lineage as players.

foremost in human organization, and the category "human being" describes one level of biotic organization that is mesa.

We can also now introduce the authors' concepts of a base objective and mesa objective. The base objective is the "criterion the base optimizer was using to select between different possible systems"; the mesa-objective is "whatever criterion the mesa optimizer is using to select between different possible outputs." The principle of selection; in other words, the metric or letter of assessment.

Crucially, while the authors discuss systems which are two- or three-level, to be a mesa optimizer is merely to stand in relation to another level. It is not an objective and inherent property of an optimizing system, but the situation of being embedded within another optimizer. This brings us to nesting and hierarchy.

Let us take seriously some form, conforming however closely as is needed for the case at hand, of Karl Friston's theory of Markov blankets. This theory holds, among other things, that boundaries are a precondition of life itself (and of complexity more generally). They are a prerequisite for maintaining homeostasis, that is, to control and regulate internal conditions which are, again, necessary to fulfilling its goals. In other words, boundaries are, first and foremost, a selection mechanism, with both a schema for admission (i.e. an entrance game) as well as physical capacities for enforcing this preferential schema.[30] They allow valuable resources—that is, goal-furthering materials, such as water to body cells,

30 We call "Trojan horses" those agents or strategies which (1) mimic goal-furthering resources in order to win entrance games and gain access to the internals of a wrapping superorganism; (2) upon admission, act in a self-interested way that runs counter to the goals of their host.

income to institutions, or weapons to a fortress—to enter and remain inside the boundaries, to assist the bounded entity in its goals. They keep undesired or harmful materials outside, either by preventing entry or expelling them. This includes not just materials and resources but also other agents or subagents, each of whom will attempt to improve its own lot by gaining access to the internally hoarded resources of another bounded agent—either antagonistically, through theft or violence or deception, or cooperatively, in symbiosis. (In reality, this is not so much an "either" case, as it is "a bit of both"; strategies are mixed, and even the notorious cordyceps fungus spans a continuum from parasitic to symbiotic.) Alignment is the central principle which separates good from bad, desired from undesired, to a mesa optimizer: it is the property of furthering or thwarting the blanket's goals. But because "goal-furtheringness" as a property of an assessed agent is a prediction about that agent's behavior in future situations, the relevant entrance games are always speculative, necessarily time-surrogative and optikratic; auto-immune disorders are one example of the failure of such assessment systems.

Our target domain, in understanding formal surrogation, is the alignment of mesa optimizers to their base optimizers, from both the perspective of the mesa optimizer and the perspective of the selecting base. Life shows a "propensity," Friston et al write, "to form multi-level and multi-scale structures of structures": hierarchy is nesting or "wrapping" layers; each layer is coordinated by one or more internal games of preferential treatment, which naturally necessitates a sort of incentive structure for interested players. Each layer stands as a base optimizer to the level below it, which is a mesa optimizer from the perspective of the layer above

it. Each wrapping layer attempts to align the goals of the blanket below it, setting the game rules of participation and preferential treatment by which complex coordination is achieved. We might call this practice "management."

For instance—and this elides necessary nuance for the purpose and pattern-emphasis[31]—a company selects employees it believes will further its goals. Alternatively phrased, people select people who they believe will further their own goals. The hiring board may pay lip service to servicing the best interests of the organization, by selecting only the "best" candidates, but this is a short-hand which ignores that members of such a board are themselves supervised, and can ostensibly be replaced, lose power, etc. They will no doubt have their own priorities, values, and goals (the mesa-layer to the company's base) including advancing social connections (nepotism) or altering the organization's composition to advance larger social agendas (e.g. diversity quotas, public good). But each employee is a Markov blanket in his own, composed of organs which are composed of cells; these too are selectively killed, expelled, or directed to the bloodstream depending on a similar appraisal of goal-alignment. (Perceived-as-symbiotic bacteria remain; perceived-as-adversarial bacteria are hunted by the immune system.) Above the company is a government, whose aims stand surrogate for the good of the nation; this government writes policy which selects for and encourages business practices that are aligned with national interests, while penalizing practices that are in misalignment. These governments are competing in a natural selection-style base layer that is geopolitics. (Payoffs, as in natural selection, are largely automatic or intrinsic, to use

31 For instance, the company does not choose hires, or selectively promote—it is other mesa optimizing employees of the company which do so.

Goffman's term; inter-national governance is a relatively unusual situation.) In all cases, we can readily furnish many examples by which the selected mesa optimizers' interests actively diverge from the base optimizing layer's interests, despite appearing, at first or externally, to align. We will call this *deceptive alignment*,[32] and note that—just as individuals are incentivized to feign cooperation while free-riding, mesa optimizers are incentivized to feign alignment if it is in their interests—if they can gain resources, or prevent persecution, from the wrapping base. (And there is almost always a payoff for appearing aligned; this is how alignment is brought about to begin with; see §6 Evolutions, p. 169)

Of course, employees also select companies just as companies select employees, in what are known as matching games. And romantic alignment (dating as as extended period of gathering evidence about a prospective partner's goals and their synergy with one's own), while varying by cultural and historical contingency in the extent to which males and females act as "gatekeepers" versus "applicants." This alignment is necessary if one will let another entity past or into one's own boundary, one's own blanket—not just exposing one's underbelly but one's tender interior. So we keep door policies, a drawbridge and portcullis.

5.7. PORTING & INDEXICALITY

Implicit in the surrogation frame, thus far, is the idea that surrogates are a sort of heuristic. They make judgment and

32 Strictly speaking, perfect alignment is impossible between two different agents, but we will speak grossly in terms of producing more value for one's ally than one seizes or detracts.

assessment easier ("cheaper") while variably increasing these assessments' error rates. In some cases, they are necessary and unavoidable; in others, they are voluntarily taken on to save compute or objectivize assessment.

Perhaps the defining feature of a heuristic is that it is contextual and contingent—what I'll call "indexical," following the ethnomethodological tradition. It saves time, or compute, only given certain environments or inputs. It relies on simplifying assumptions which may prove unwarranted, if lifted outside its home environment.

That is, surrogates in full-bodied selection games (games between two or more full-bodied agents) are scoped to an expected set of player strategies, constraints, and backdrops. When there are only two cases in need of distinction, an assessor can isolate a few, or even a single, criteria of difference by which to distinguish the cases, for instance, distinguishing gold from pyrite by the color of streak it leaves behind on unglazed porcelain.[33] We can understand this partially through information theory: as Simon DeDeo writes, in a game like "20 Questions," there are better and worse questions for eliminating possibilities and identifying the object the question answerer holds in mind. DeDeo refers to these strategies are more or less optimal given a specific opponent and the distribution (kinds and their relative frequencies) of objects this opponent is likely to keep in mind. It is equally true that one could come up with a set of optimal questions, or an optimal questioning strategy (strategy of distinction) across all possible opponents, but crucially, this strategy would be less fit than such a strategy tailored to a given opponent.

33 This is known as a streak test; gold leaves a yellowish streak, while pyrite's streak is a greenish black.

Moreover, this optimality would still be highly contextual, would be bound up with the structure of existing things. That is, if the questioner has already determined that the object-in-mind is a mammal, it may be efficient to ask whether most cases of this mammal are domesticated or under human supervision and breeding regimes. But the incisiveness of such a question, which distinguishes between wild and domestic mammals, only works because of the actual landscape of mammal life at a given place and moment in time. If virtually all the mammals on the planet, or more accurately which a question answerer is likely to have in mind, were wild, then very little information would be gleaned by such an inquiry. It is precisely because approximately even proportions of mammals, which we would imagine an average answerer to hold in mind, are wild vs. domesticated, that such a question is efficient at distinguishing.

But when possible not-golds are nearly infinite, or the weighted distribution of possibles unknown, there is no reliable method of distinguishing other than the full evaluation of each assessed object's property in its entirety. Many surrogate markers work efficiently because they more or less accurately carve a set of common cases or "moves" that a selected party is expected to make in trying to win a selection game.[34] This leads to selection pressure on the surrogate that alters the composition of the set of common moves, defanging the surrogate. Nassim Taleb advocates, in his *Incerto* series, the heuristic of preferring, all else equal, a surgeon who is slovenly and interpersonally bracing—the idea being that, if he

34 Moreover, implementing multiple distinguishing tests is costly, requiring the expenditure of time and other resources by both the evaluator and the evaluated. And since many such selection games are "matching games," selectors may decrease the number of tests they run on potential candidates in order to make themselves more attractive to said candidates.

has managed to attain equal rank, at his hospital, as a more charismatic, well-kempt surgeon, then he must be a better surgeon. This is not bad advice on the face of it; it is merely short-term advice, as its usage undermines its own efficacy; its efficacy is reliant on the dominant incentive structure (the dominant system of surrogates) disproportionately selecting for the charismatic and professionally attired.[35]

When the environment changes—when new cultural, expressive, or literal technologies emerge—the space of possible, likely, known, and frequently employed player strategies changes. And this makes a given surrogate less "fit" as a heuristic for distinguishing between them. This fit between surrogate and assessed party (and assessment domain) is central to the quality of the surrogate as surrogate.

Similarly, the porting of surrogates across domains, cultures, and game tournaments ought to be undertaken only with extreme care. The imposition of evaluative systems that make sense given one set of assumptions will break down when subjected to a very different set of behaviors.

35 Elsewhere, Taleb argues that it is a fallacy to believe an attractive apple is a good apple (in the visual vs. taste sense). This position, too, deserves a fair bit of contextualization and nuancing. Animal visual perception evolved to detect quality fruit, just as fruit-bearing plants have evolved to visually signal taste and nutritional content. (Taste being a second level of evolved surrogate for nutritional value.) That is, in general there is a tight, evolved relationship between quality and visual appearance—we really *should* judge this book by its cover—a point which becomes obvious when we consider rotting and desiccated fruit. It is specifically on the margins of high-quality fruit that looks and taste can uncouple, as genetic engineers or breeders select for one at the cost of the other. In other words, the surrogate falls apart (becomes unfit) under selection pressure—at least in the short-term.

In adversarial games, players are incentivized to push opponents into environmental spaces and problem distributions where their surrogates become uncoupled, their heuristics for perception and action less fit. A strategically naive player, observing that a given game tactic is only rarely employed in contests, might forego investing in counters to said tactic, figuring that he can afford to forefeit the occasional point lost to it. He will quickly find that the "rare" tactic is now used constantly and unceasingly against him—in other words, that its relative rarity was purely a consequence of players' historic investment in an arsenal of counters.

This advantages players with faster, more flexible OODA loops (i.e. players who are more adaptive and generally intelligent), as they are capable of pushing ("treadmilling") the meta into new positions, and seizing emergent arbitrage opportunities, faster than their opponents can adjust.

Human general intelligence therefore represents one of the greatest strategic advances in evolutionary history. Consider, for instance, the sort of stupid insistence—the lack of adaptation and context-fittedness—which we see in animal signaling studies, when test subjects are exposed to game states they did not "train on" (evolve to master):

> Tinbergen found that birds would sit on an oversized plaster egg with heavily defined markings and saturated colors, preferring this supercharged model to their own pale, dappled eggs. Male butterflies would attempt to mate with gaudily painted cardboard dummies in preference to real females. Gull chicks would attempt to feed from a red-striped vertical dowel, ignoring their parents' beaks to the point of starvation. Geese would ignore their own eggs and tirelessly strive to roll

a volleyball into their nest if it was adorned with the appropriate markings. A stickleback would attack a painted wooden model in order to defend its nesting territory, so long as the model had a schematic "eye" and a red underside.

Although [the organisms] are triggered by stimuli that are indicators of healthy, vitality, danger, or reproductive advantage, the unlearned actions of animals respond not to an exhaustive assessment as to whether an object satisfies these criteria, but to abstracted perceptual cues (blueness, egg size and shape, red underbelly) with which those assets or threats were consistently associated over the evolutionary period.

5.8. THE TYRANNY OF ROUND NUMBERS

In bargains that involve numerical magnitudes, for example, there seems to be a strong magnetism in mathematical simplicity... [a] tendency for the outcomes to be expressed in 'round numbers]... [or] the frequency with which final agreement is precipitated by an offer to 'split the difference'."

Schelling, *Strategy of Conflict*

At *The Conversable Economist*, Tim Taylor reports[36] that consumers often stop filling their gas tanks at rounded dollar amounts, or will give even-dollar tips to waiters; that baseball

36 "Round Number Bias" 2013.

coaches make decisions about players based on round-number cutoffs:

> [Pope and Simonsohn] find, for example, that if you look at the batting averages of baseball players five days before the end of the season, you will see that the distribution over .298, .299, .300, and .301 is essentially even—as one would expect it to be by chance. However, at the end of the season, the share of players who hit .300 or .301 was more than double the proportion who hit .299 or .298. What happens in those last five days?
>
> They argue that batters already hitting .300 or .301 are more likely to get a day off, or to be pinch-hit for, rather than risk dropping below the round number. Conversely, those just below .300 may get some extra at-bats, or be matched against a pitcher where they are more likely to have success. Pope and Simonsohn also find that those who take the SAT test and end up with a score just below a round number—like 990 or 1090 on what used to be a 1600-point scale—are much more likely to retake the test than those who score a round number or just above.

The interest in round numbers arguably ought not to be considered a cognitive bias, since those who optimize for round numbers are intelligently adapting to the round number "biases"[37] of others. (Similarly, narcissism may fairly be

37 "Biases" in scare quotes—there is a compelling argument to be made, which is out of scope for these pages, that round numbers act as Schelling points, i.e. natural resting points for coordination, and that parties' outsized interest in round numbers are at least partially (and rationally) premised on this fact.

 considered as a strategic response to an optikratic society, rather than a pathology.)

Readers, no doubt, are well-familiar with the practice of selling products for some variation of $X.99, so as to not advertise a price of X+1. Studies support the intuitive finding that round decade-markers (1960s, 1970s, etc) disproportionately influence our understanding of personal and cultural histories, and that decade birthdays (30, 40, 50, etc) are accompanied by significantly more emotional angst and "meaning crises" than non-decade birthdays. Sam Fussell, in 1989's *Muscle*, writes of the weight-lifting magazines he perused in his twenties: "From what I could glean from the magazines, real builders, like Arnold, Bill Pearl, Lou Ferrigno—all of them had 20-inch necks, calves, and arms; 30-inch thighs; 60-inch chests." Round numbers become influential in sports punditry and fan followings; for decades prior to the first recorded sub-four minute mile, many believed the it an unbreakable barrier, though why four minutes instead of 3:59 or 4:01 defined the point of human impossibility was never fully established. In the NBA, Russell Westbrook's triple-double average in the 2016-17 regular season won him an MVP award; falling slightly short of double digits across four categories rather than three is a more impressive but less-lauded athletic achievement, as it "slips through" the base-ten system of performance-tracking that the Association and sports fans have crystallized.

Similarly, the incentivization of round numbers means optimization towards them. Gioia & Corley, in 2002's "Being Good vs. Looking Good: The Circean Transformation From Substance to Image," chronicle that, after the advent of business school rankings starting in 1988, "significant changes have attended every business school that aspires to

be declared a 'top-10, top-20, top-40, or top-50' school by these two progenitors [*Business Week* and *U.S. News*] or their many subsequent emulators."[38] One can imagine a difference, in dating prospects (outcomes), as well as online profile number-fudging (behavior reflecting an implicit belief in disparate outcomes), between two men each of 70.5 inches height—one, American, whose height is measured in feet; the other, European, whose height is measured in meters. (A minimum height of six feet is a common cutoff in selection assessments of men, which has been exacerbated by online dating featuring individual "statistics.") This difference has nothing to do with the two men's "natural" fitness, but rather their sexual fitness *within a formal system* created by the culture they find themselves within.[39] Naturally, evolutionary selection has become deeply interwoven with cultural process, as so many of its surrogates are culturally defined.

This is only to illustrate the degree to which neutral, even banal, features of the cartographic or "reductive" system

38 The University Affairs job listing site included, as of October 2021, an open position as University Ranking Strategist, offering a pay range of $81-135k. See Sauder & Espeland's Engines of Anxiety for a full treatment of competition around law school rankings as a surrogate for university prestige (driving e.g. donations, applicant quality, and alumni hireability).

39 Formal systems aside, male height is an interesting surrogate for fitness because, while in an ancestral environment height is intrinsically important for selection games (such as those against large cats, or physical altercations against other tribes), in modernity it is more or less irrelevant intrinsically. That is, its entire value is speculative and *extrinsic*, founded on our own vestigial, surrogative prejudices from politics to salary comp. The surrogate has been reified even as the environmental context in which it predicted intrinsic fitness has disappeared. And as in Keynesian beauty contests, the Matthew effect, Tinkerbell effects, and runaway sexual selection, a fetish for height leads others to speculatively prioritize for height, insofar as taller children will themselves be sexually preferred, and so on.

can lead to substantial changes in human behavior. Similar examples can be found not just with round numbers and base-ten variants, but in numbers like three, five, and twelve. Analog gradients need compressing and systemization; this is the very function of structures, including language, as Nietzsche details in "Truth & Lies"; this compression makes it computationally *tractable*, which is to say, able to be handled and manipulated. But there is a dark side to this miracle, which is its vulnerability to gaming.

6. Evolutions

6.1. ECOLOGICAL PERSPECTIVES

Let us switch perspectives as we conclude this conceptual exploration, and make good on the evolutionary potential implicit in a framework like the selection game.

Perhaps the most crucial foundation of the surrogation concept is our situation, as organisms, of ecological interdependence, where each (ecologically huddled) organism's decisions affect each other (ecologically huddled) organism's situation. Within the huddle, organisms' actions and states are partially observable to other organisms in the huddle.[1] In the struggle for life and reproduction, organisms learn to predict—to interpret, and "read"—one another in order to optimize around one another—that is, around their dynamic environment. This ability to read gives way to the ability to "write"[2]—to act in a way that, when read, influences the reading organism in a way advantageous to the writing

1 In ecologies (like the city) which approach the quality of a rainforest, player environments are dominated by the strategies and actions of fellow players; games tend toward extrinsic (in the Goffmanian sense). Agents and strategies that persevere in such an environment, either through agent learning or natural selection, will end up fitted to one another. Relatively asocial "desert" environments, on the other hand, will be more intrinsic. The Inexact Sciences blogger Feast of Assumption (2022) has similarly contrasted the dynamics of "PvP" (player versus player) with "PvE" (player versus environment) games.

2 This "generalized reading" idea—that the basic literary-theoretic process is emblematic of ecological interdependence and by extension, the social world—is described at greater length in a series of letters "On Generalized Reading," at Letter.Wiki.

organism's survival and reproductive chances. (Or, perhaps, for the great surrogate of evolutionary fitness: pleasure.)

What we desire, in the abstract, is more or less constant, and difficult to influence or manipulate. But what we pursue concretely, as instantiations or means of securing our abstract desires, depends on perceptions, interpretations, beliefs, self-concepts (about the environment, about our own wanting). That is, the concretia of decisions are mediated by organisms' more readily manipulable epistemological practices—their sense of what is, their sense of causes and effects, their sense of the possible and probable. We have no access to the facts—only guesses, informed by physical clues in their environment which we read and organize into patterns in a process C. S. Peirce called abduction. Manipulation of these testimonies, in the service of manipulating the reading organism's behavior, is one of the most powerful tools organisms have for securing their own self-interests. Thereby do mimics carve out a niche: by evolving or designing testimonies that manipulate the predictions, and therefore actions (to eat or not to eat) of ecological co-inhabitants (e.g. predators). By *writing*.

An incentive structure, or internal game, is an artificial module of additional reward functions, whose dispensation is socially mediated, which is, typically—although not always, as in the case of citizenship—voluntarily joined[3] (pending the

3 This voluntary participation often comes at the formalized or functional exclusion of some other incentive structure (game, payoff matrix, reward function, etc.). I say "functional" exclusion because there is either a physical exclusivity to achieving goals in two given games simultaneously, or because, if the "multi-gaming" is discovered, the player will be ejected from the incentive structure of a given gaming module (as in the case of a double agent, or an individual who writes publicly about his institution. Nightjack, who we will

outcome of an entrance game gating admission). One player, or a set of allied players, offer rewards for certain kinds of behavior which promote their (that is, the hosts') interests. This structure is, essentially, the manipulation of concrete agent goals not through epistemic manipulation—appearing to be something the agent desires, or does not desire, as is the case in a selection game—but through the creation of a system for dispensing rewards conditional on behavior which benefits the creator or host of the reward function. (It is somewhat like a class extension in programming: the original reward functions of the world remain in place, but an additional module is added which offers the concretia of human desire—at a cost.) In attempting this manipulation or recruitment of other ecologically proximate players through the lure of reward, the surrogation problem is introduced. It is appearance of meeting the game criteria which secures the dispensation of reward, and thus the signs of meeting the game criteria are optimized for by players.

And since the rewards of an incentive structure are social contracts extrinsically enforced—the pleasure of a peach is intrinsic; it happens automatically; but a paycheck must be mailed—their dispensation is the result of judges reading player performance through these same manipulable signs (surrogates). In such situations of extrinsic reward allocation, and in a reward module where the benefits sought by agents come at a conditional cost, then outsized benefits can

---

cite shortly, served as a police officer while writing an anonymous police blog which eventually won the Orwell Prize. When his identity was "doxxed" by the UK newspaper *The Times*, in the wake of the prize, he was compelled by his CO—perhaps on order handed down from up high—to remove his blog and cease writing. Many similar stories abound of academics in the 1990s and 2000s, who faced formal sanctions, or were merely disadvantage in institutional selection games, by their blogging.

be gained at a proportionally less cost by manipulating the signs of compliance (the signs of payment). These surrogates must—like all effective currency—therefore be difficult to fake, and easy to verify (these being flip-sides of the same coin, from the perspective of the selected and the perspective of the selector, respectively).

When the surrogates used by such an institution for dispensing rewards, and evaluating work or contribution the institution's external game, are "high entropy," they are prone to degeneration; the institution is overtaken by individuals who are not aligned to advance the institution's external goals, and the institution eventually falls apart.

6.2. FEEDBACK LOOPS

Surrogation may not appear so serious a problem unless one considers that many games are iterated and evolutionary. That which is selected for perseveres; given enough time, those with even slight advantages will outcompete the rest. And in many selection systems—for instance, institutional advancement and promotion—those who win at lower-level games become the designers and enforcers of higher-level games, influencing the selection process itself.

Insofar as an institution is a body of individuals, with varying capacities as decision-makers, varying ideals of integrity, communicative capacity, and coordinative inclination, it is the selection game—the assessment which qualifies an outsider to serve within an institution, or an insider to climb the ranks of power—that counts most. Selection games are the screening mechanism which keeps eccentric talent out, or mistakes glittering image for actuality; constructs a cycle of

accreditation, or a pseudoscience out of psychology. Rules and culture, the "structure" which is more regularly blamed for the shortcomings of bureaucracy, are determined by early members, even they are merely an influential byproduct of the org's first selection games.

We've already discussed how an ideal of transparency can undermine the efficacy of even good-faith assessment, by making the surrogative basis for selection known and thus more easily gamed. A fake cannot be like the original in every way without becoming genuine in its own right.[4] But assessment surrogates are necessarily partial, such that perhaps only a few axes or properties (e.g. drug dogs and x-rays for border control, or streak tests distinguishing gold and pyrite) need to be beaten or faked. Those gaming the system can focus resources and energy on the narrow criteria.[5] But as we have also seen with the sugar maple or Vavilovian mimicry examples (§1.1 Single-Agent Selection, p. 13), learning across a population occurs evolutionarily; indeed,

4 E.g. rye's transition from weed to cereal.

5 Daniel Dennet, in *Consciousness Explained*, writes: "Hallucinators usually just stand and marvel. Typically, they feel no desire to probe, challenge, or query, and take no steps to interact with the apparitions. It is likely... that this passivity is not an inessential feature of hallucination but a necessary precondition for any moderately detailed and sustained hallucination to occur... The reason... hallucinations can survive is that the illusionist—meaning, by that, whatever it is that produces the hallucination—can "count on" a particular line of exploration by the victim... So long as the illusionist can predict in detail the line of exploration actually to be taken, it only has to prepare for the illusion to be sustained "in the directions that the victim will look." Cinema set designers insist on knowing the location of the camera in advance—or if it is not going to be stationary, its exact trajectory and angle—for then they have to prepare only enough material to cover the perspectives actually taken... In real life the same principle was used by Potempkin to economize on the show villages to be reviewed by Catherine the Great; her itinerary had to be ironclad."

natural selection is the archetypal example of a selection game, and its theoretical origins lie in breeding and domestication. There is an implicit, evolutionarily knowledge of the selection game surrogates encoded in the maple tree or rye DNA.

In *The Wire*—perhaps the great artistic depiction of institutional surrogation—we see how more extreme—and explicit—examples of corruption feed-back loops occur. "Straight" cops are more difficult to manage and manipulate to corrupt ends than "bent" cops, and thus a bent officer with hiring and firing, promotion and demotion power will prefer other bent cops. Cops who are straight will be incentivized to bend, and those who are already bent will be preferentially selected (to become future selectors). Some of these dynamics have come to light in recent coverage of Los Angeles County Sherriff gangs, and corruption more generally across the L.A. prison system:

> Long before Tanaka officially inherited the No. 2 spot there were already two camps inside the Sheriff's Department—those "in the car" with Tanaka and those on the outside. Those outside the car can be "rolled up"—meaning transferred to department backwaters—if they cross Tanaka, regardless of their performance on the job. Those in the car with Tanaka are promoted quickly and insulated from performance failures...
>
> Tanaka gives out [loyalty] coins to only a selected few, and each coin is serially numbered, in part, so no forgeries can be made, but mostly to emphasize the special nature of the talismans. They are earned, say sources, through loyalty to Paul Tanaka. "I can't prove it, but

> from what I've observed, there are two ways to get ahead in this department," says retired LASD commander Bob Olmsted. "The official way is the civil service way of solid performance reviews, expected performance and various forms of testing. The real way is to become a 'Tanaka boy'—by volunteering and donating to his campaign and smoking cigars with his inner circle."[6]

In many cases, such as in academia or journalism, the feedback loops are less overtly corrupt. Institutional states typically attributed to conspiracy, such as the political biases of major media and scholarly organizations, are more frequently the result of tacit selection. Editors do not mandate, top-down, the political slants of their network, but shape it bottom-up through hiring decisions, story selection, applicant self-selection (the institution develops a reputation) etc. These decisions are in turn made predictively with respect to expected value based on audience selection—the kind of stories that are widely read and shared, etc—as discussed in §1.4 Institutional Nesting, p. 24.

In inexact science fields like psychology, as a result of surrogation (operationalizing abstract nouns like "anger," then naively treating these surrogates as adequate), researchers "expend enormous resources on studies that are likely to have very little informational value even in cases where results can be consistently replicated."[7] Statistically and inferentially unfounded claims are passed up, from inexact

6 Fleischer, "Dangerous Jails" 2013. Robert Jackall, writing in *Moral Mazes*, presents a similar case: "Younger managers learn quickly that, whatever the public protestations to the contrary, bosses generally want pliable and agreeable subordinates, especially during periods of crisis. Clique leaders want dependable, loyal allies."

7 Yarkoni 2019.

science labs, 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 psychologists who endorse generalities that their studies do not support. There is widespread Goodhart-style gaming of statistics of legitimation, the most well-known being p-hacking. Yarkoni presents a number of "next steps," given this 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 sociological angle from whence such 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.[8]

The first problem is that more modest claims come at the loss of power, prestige, and reputation. Not only would fields and institutions investigating inexact science issues be ceding their previously claimed credibility, but any individual researcher making more modest claims would be outcompeted in receiving grants, public speaking and policy consultation opportunities, etc.

The second problem is that as individual researchers 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. There is a demand for operable, generalizable social and

8 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.

psychological insight which requires only some researchers to supply it. Replacements will, on average, have less integrity, less rigor, and less knowledge as to the limitations of their practices than those who they replace. They will then train future students in their techniques and philosophies of science.

In other words, as knowledgeable insiders slowly leave these fields, or opt not to join their ranks in the first place, they may become increasingly destructive and ill-founded until their public credibility begins to collapse. This process has been with inexact fields 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 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 obstacle.

Inexact scientists 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.[9] Misuse raises the bar of expectation; those who optimize toward legitimate scientific practice—in other words, understanding 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

9 From the perspective of the employed player, being fired is equivalent to death—the end of play, ejection from the game.

not just the career prospect of individuals but the larger efficacy and service of the institution.

Many angles are taken in analyses of institutional failings—conformity, risk-aversion, asymmetrical justice, preference falsification. But the changing nature of its selection games—the ritualization of their spirit, the delegation of its oversight to successive generations of HR and managers, and the feedback loops of selection and company culture—is often overlooked.[10]

6.3. FADS & ANTI-INDUCTIVITY

For each selection game, there objectively exists a set of solutions: possible courses of play, or combinations of moves, which ensure a given outcome. Some of these solutions are intended, or "designed": there is a "right" way to gain acceptance to Oxford, and ways which, if discovered, would bring official or social sanction. We will call the former cooperative solutions, since they tend to benefit both the selector and selected, and obey the spirit of the game. The latter, meanwhile, tend to be adversarial or parasitic, in that they usually benefit the solution finder at the cost of the selector.[11] Part of the pleasure of films such as the *Ocean's* series is watching

10 The lemon problem in economics is one area in which these feedback loops have not been ignored.

11 This is of course a simplification, and not necessarily the case: we may deceive for the benefit of the deceived party; the assessor may not act in his own interest; honest cooperation and beneficial outcomes are distinct axes. But premised on the assumption that actors tend to broadly be aware of their own interests, and tend to be roughly competent at running selection games which advance them, there will be a resulting correlation between "playing the right way" and advancing the assessing organism's intersts.

a group of individuals solve an expensive and elaborate selection mechanism designed to only give a "select" group of individuals access to the casino's inner sanctuary (and by extension, its money supply).[12] Every castle can be penetrated, as Homer's Trojans teach us—if only Cassandra had won the selection game to have her advice heeded.

Because the criteria in a selection game are often designed to be both opaque and robust to adversarial exploitation—as in the *Ocean's* casinoes, or the prescribing of amphetamines or opioids to patients—it can often be difficult to stumble upon one of the available solution routes. If the criteria are not publicly available, trial and error may be necessary. But once a solution is stumbled upon—or devised, tested, and proven—it will quickly spread, *provided that* the solution has a means of reaching more public awareness. Because exploits, once known to the selection game designer or host, can often be effectively patched, it frequently behooves those who solve a selection game to conceal their solution. This can be difficult because merely by using a solution, a player can leak information to other players including the selector himself. Card-counters in Las Vegas, should they beat the house more often than they ought to, will be banned from the casino. In poker, a player who discovers an opponent's tell may purposefully lose certain small-stakes rounds, to prevent his opponent from realizing that the tell has been discovered (and thereby strategically using it for deceptive purposes, consciously displaying it when one is in fact not bluffing...). In Paul Thomas Anderson's masterpiece *Punch-Drunk Love*,

12 The casino, of course, has been selected by the writers because its funds will not be seen as "honest" money—its games are "rigged" against players, in favor of the house—which allows the audience to cheer on the heist perps in good conscience.

the protagonist Barry Egan keeps hush-hush a discovered exploit to an American Airlines frequent flier program, out of fear that, should it be exploited by others first, the airline will move to discontinue the program. That is, merely by someone cashing in on the miles degenerately, the loophole may be closed.

As one pseudonymous doctor writes of his time spent in Haiti,

> [Haitians have the mindset that] getting more medicine of any type is always a good thing and will make them healthier, and doctors are these strange heartless people who will prevent them from taking a stomach medication just because maybe they don't have a stomach problem at this exact moment. As a result, they lie like heck. I didn't realize exactly how much they were lying until I heard the story, now a legend at our clinic, of the man who came in complaining of vaginal discharge. He had heard some woman come in complaining of vaginal discharge and get lots of medication for it, so he figured he should try his luck with the same. And this wasn't an isolated incident, either. Complaints will go in "fads," so that if a guy comes in complaining of ear pain and gets lots of medicine, on his way out he'll mention it to the other patients in line and they'll all mention ear pain too—or so the translators and veteran staff have told me.[13]

Another way of stating this is to say: Actors in an optikratic landscape are constantly watchful as to the significance of—which is to say, the structure of payoffs accorded to—different actions, cues, and appearances. Since most of humans'

13 LiveJournal 2011.

games are fundamentally social, humans' assessments today deeply structure the kinds of optimizations that will be implemented or evaluated tomorrow.

We will call what results a *solution fad* or a *solution cascade.*[14] Perhaps most enticing is the possibility that all fads are solution fads. When a solution to a selection game is discovered and widely adopted, it inevitably leads to a new equilibrium of play. Information leakage leads to wider discovery and adoption: by playing a card, one cannot help but show the card. We will call such situations *anti-inductive*. There is no stable strategy, because play can never be globally optimal, merely optimal relative to other player strategies. What's more, any new solution or strategy, on use, becomes available to other players for adoption. In buying stock or choosing one's fashion statement in the morning, one reveal's one's strategy, and makes it available to mimicry. Today's matrix of visible payouts is tomorrow's set of symbolic performances.

Solution fads do not merely happen in Haiti, among the under-educated; they are equally characteristic of the Western legal system. Nightjack, an anonymous police blogger who won an Orwell Prize in 2009, writes in his 2008 entry appropriately titled with a sports metaphor, "Goalposts Moving":

> PC Ellie Bloggs posted on her blog that manslaughter is the new murder. I have to take slight issue with that. Manslaughter is still the old murder, it is just that now

14 This term is inspired by Timur Kuran's *preference cascade*, in which a previously suppressed belief (suppressed within a "preference regime") is increasingly vocalized. Each vocalization makes subsequent vocalizations politically safer, creating a positive feedback loop and rapid, widespread adoption of a belief once it is articulated by some small but critical mass. (The Emperor's New Clothes being a mythic telling of this sociological phenomenon.)

> it has a smart lawyer and a psychiatric report. I speak in terms of the gradual extension of the doctrine of diminished responsibility. This is a defence that can reduce murder to manslaughter. Was it a new law? Nope, just the usual judges having another look at where the boundaries should be placed...
>
> [A]ny sensible defence lawyer will be looking for a psychiatric report containing such phrases as "Adjustment Disorder...," "Personality Disorder...," "Severe Personality Disorder...," "Depressed," "Morbid Jealousy," "Post Traumatic Stress Disorder," "Persecutory Delusional Disorder," "Alcohol Dependency Syndrome...," "Acute Stress Reaction," "Khat/Amphetamine Psychosis." [...]
>
> Once a defence team hear that somebody has gotten one of these [defenses] home, strangely more defendants seem to start suffering from it. It's a bit like nationality/religion/persecution stories with asylum seekers, where the circumstances leading to a successful application become viral.

This is not to say such pleas are always fabrications. It is to say that, unless one's reality happens to fall into an established category (that is, a decision rule), then one must fabricate in the direction of such diagnoses, in order to compete and be heard within such a system. It is to say that many who submit such pleas are submitting them primarily because they are tactics known to work—that terms like PTSD are surrogate markers in a high-stakes selection game, and therefore act as behavioral attractors.

As fads emerge, evaluators catch on and begin devaluing the faddish cue, since it is being widely free-ridden by those who lack the qualities it implies (but wish to appear as if they possess them). We can see this in job applications, college admissions, and ADHD prescription requests—individuals looking to pass the requisite selection test spread word of successful entrance strategies. Inevitably, the Red Queen has her way; the newly found "solution" no longer works, and a new "passcode" or set of "magic words" that open the selection gate must be found. There is a "metonym treadmill" by which players are constantly seeking metonyms to gain entry, and gatekeepers are constantly seeking to keep their metonymic interpretations accurate, to prevent their being "hacked." Those who follow a fad, such as advancing false claims of mental illness in court, actively hurt the long-term prospects of the actually ill.

One advantage of informal evaluation is that its saturation sensitivity is continuous—that is, individuals can have a rough sense of the population frequency of a certain solution, and apply an inflation-style penalty by devaluing the solution.[15] Formal games, on the other hand, can only deal with saturation discretely, by writing new laws. There is the period before the law is passed, during which the solution has "full value," and a period after the law is passed, when the solution has been discretely de-valued (either penalized, to compensate for its advantage, or outright banned). There

15 The current surrogation regimes common of formal, institutional selection are primarily entropic, as a result of the positive feedback loops acting on and increasingly the level of soft corruption, and due to a lack of necessary adaptive work by selectors. A healthy institution would have to solve the problem of not being taken over by degenerate players (and thereby degenerating...) which requires negative feedback. In this way, institutions could possibly learn something from, say, the fashion landscape.

is inevitably some tipping point of adoption which triggers this update to the internal game's letter, which in Major League Baseball has been the rapid, statistically significant decrease in batting percentages as "sticky stuff" is widely adopted by pitchers across the league, allowing them to throw more difficult-to-hit pitches:

> "Pitchers are shortsighted if they're not mad [about sticky stuff]," says Marlins reliever Richard Bleier, who says he has never used anything more than sunscreen and rosin because he wants to feel proud of his career. "Like, 'Oh, we don't want hitters to hit'—well, look what's happening now. Hitters aren't hitting, and now everybody's going to be penalized."

Of course, this process of adjustment can be painful; economist Eric Falkenstein speculates that a good deal of economic boom-and-bust cycles are the result of surrogation problems, which he likens to Batesian mimicry:

> In an expansion investors are constantly looking for better places to invest their capital, while entrepreneurs are always overconfident, hoping to get capital to fund their restless ambition. Sometimes, the investors (dupes) think a certain set of key characteristics are sufficient statistics of a quality investment because historically they were. Mimic entrepreneurs seize upon these key characteristics that will allow them to garner funds from the duped investors... The mimicry itself may involve conscious fraud, or it may be more benign, such as naïve hope that they will learn what works once they get their funding, or sincere delusion that the characteristics are the essence of the seemingly promising activity... Once the number of mimics is sufficiently high, their valueless

enterprises become too conspicuous and they no longer pass off as legitimate investments. Failures caused by insufficient cash create a tipping point, notifying investors that some of their material assumptions were vastly incorrect.[16]

6.4. EXPRESSIVE TECHNOLOGIES

One way to understand these solutions is as *expressive technologies*. We could equally call them *impression technologies*, insofar as "expressive" perhaps implies, under the reigning ideology of late 20th and early 21st century English-speaking culture, a Whitmanian self-expressivity, the truthful public articulation of interiority. Nothing could be further from the

16 Falkenstein 2010. Note that the dupes in question are dupes precisely because they fail to recognize the anti-inductivity of the game they are playing. This leads them to take a simplistic stance on historical data, and to rely on statistical analyses as if the problem were simply inductive. Falkenstein continues:

> The key is that the mimics and duped investors chose those business models that seemed most solid based on objective, identifiable characteristics that were, historically, correlated with success. An econometric analysis would have found these ventures a good bet, which is why investors did not thoroughly vet their business models. For example, banks stocks through 2007 were one of the best performing industries since industry data has been available in the US, and performed well in the 2001 recession.

Recall, from §5.7:

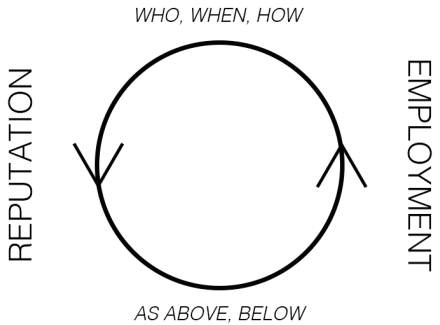
> A strategically naive player, observing that a given game tactic is only rarely employed in contests, might forego investing in counters to said tactic, figuring that he can afford to forefeit the occasional point lost to it. He will quickly find that the "rare" tactic is now used constantly and unceasingly against him—in other words, that its relative rarity was purely a consequence of players' historic investment in an arsenal of counters.

working of an expressive technology, whose task is securing a desired impression, e.g. in a selection game.

Briefly, in such a selection game, an expression wins by securing a desired outcome. An impression wins by providing an accurate algorithmic model along axes relevant to the selection system. By "algorithmic" here, I mean that it provides a useful predictive model for how the assessed subject (or object, in a single-agent selection game) will behave, in the ways which are pertinent to the project and goals of the selector.

An expressive technology is a symbol, speech act, framing, metaphor, implication—more or less, a vector of communication—that is employed with the goal of securing a desired impression. It is the primary means by which an assessed subject manipulates their assessor's assessment. An expressive technology that "works" (secures the desired impression) is a solution fad in the making—*provided* that its use leaks information to other players, and can be widely copied. That is, if it becomes a fad, it will be a fad relative to a population of competent players who can and are incentivized to adopt it. (Some fads are "blocked" when, still in the early stages of viral spread, the larger population deploys an immune response whereby they form, spread, and associate a disgust reaction to the fad. At this point, the original fad becomes limited in its memetic spread to a subculture of players who are willing to take a social hit, and through their reciprocal social acceptance of other solution-deploying players, subsidize the solution within that subculture.)

The expressive properties (or mechanics) and the value of an expressive technology together are roughly equivalent to its reputation—the set of specific and broad impressions and regards, both at the first- and second-order. By first- and



second-order, I refer to the obvious fact that many individuals have functional disgust reactions (in the sense of behavioral avoidance and social disparagement) to subjects and objects they have never personally encountered, but which have either been learned in the abstract (first-order), or which are held in a more detached, instrumental way (second-order). One may not have any personal fear of, or animosity toward, an expressive technology, and yet understand that its employment would, if publicized, be costly. Extreme versions of this picture resemble the preference falsifiation regimes outlined by Timur Kuran.

Those technologies which would have a negative effect on a given audience's impression—that is, which would have a goal-obstructing or project-damaging effect on the player who employs it—can be said to have negtive value as an expressive technology. (Or "negative expressive value.") At the same time, these technologies are still sometimes used in contexts where they have positive *pragmatic* value. Moves

with negative net value are only ever employed out of ignorance or by accident. We can (roughly) distinguish pragmatic or "intrinsic" value from expressive or "extrinsic" value by whether an action, tactic, heuristic, etc would be worth enacting if stranded upon an uninhabited island. Hacking up phlegm has asocial or intrinsic value even as it tends to damage the impression one gives off around others. Much of human life is characterized by a behavioral divide in private versus in public, which can be modeled through recourse to private actions' negative expressive externalities.

When the expressive value of a technology is lower than its asocial value *to a given audience*—that is, when a given solution to practical problems which besiege the technology's observers is disincentivized by its negative social reputation among said observers, we can call this solution holistically underpriced. When the opposite is true, it is overpriced.

Finally, the first-order effects of an object or source are often—particularly in objects whose first-order effects are social and psychological—modified by second-order reputation (or "connotation"; in short, the technology's associative, social baggage).

This loop drives natural selection but also the social world at a much faster rate: information status and the "meaning" of symbols can change in days, hours, minutes. We saw, e.g. that in the wake of the use of the phrase "sexual preference" by a conservative judge, the definition of "sexual preference" in Merriam-Webster's online dictionary was updated within hours of the utterance to emphasize the phrase's (perceived as) disrespectful connotations.

6.5. ARBITRAGE & HETEROGENEITY

The more that a single system of surrogates (informally, a single perspective) dominates a landscape of gameplay, the more that alternate systems (perspectives) are subsidized. In more cooperative games, a heterogeneity of systems can outcompete homogeny insofar as minority or "alternate" vantage points are used to error-check dominant or "default" systems, and towards improve overall performance in those areas where the dominant system is weakest. In more adversarial games, the predomination of a given surrogate system opens up arbitrage opportunities for inventive players. Insofar as a single system of surrogates predominates as a basis for decision-making, tactical opportunities will inevitably emerge which said system fails to "see" (i.e. properly value). Marketing theorist Rory Sutherland describes such opportunities in an interview with economist Russ Roberts:

> I was asking people about this just the other day, "How should you use the London Tube Map to buy a house?" And there are two answers to it, "I want to buy a house near the Tube," or "Everybody else uses the Tube Map when deciding where to live in London. So what I've got to do is actually look at what isn't on the Tube Map." And, in many ways, if you think about it, South London—without becoming a sort of London transport bore at this point—South London's rail network is very, very well supplied with trains, none of which appear on the [Tube] Map. And you can probably buy insanely undervalued property next to a railway station south of the river, which is actually half the journey time into work versus, say Fulham, which is on the Tube. And the reason you're getting that bargain is partly because

you're using a different model of choice as everybody else. And so you're looking for what's undervalued.

6.6. CLOSE AND DISTANT EVALUATION

Daniel Boorstin 1961, *The Image*:

> In an age when the average consumer has only the vaguest notion of the actual activities of a vast, complex corporation, the public image of the corporation substitutes for more specific or more circumstantial notions of what is going on.[17]

Cristóbal Sciutto, "Lacunae, DIY, and gestalts" 2021, discussing Georg Simmel's "The Metropolis and Rural Life":

> [In the city,] personal identity can only emerge through attention-grabbing signals (e.g. public lists of books that one has read, pathetic in retrospect), yelling "I am here." [In] rural life... there is space for one's emotions and, in aggregate, personality to emanate from the self. One interacts with few people, repeatedly, for long periods of time. One's uniqueness and irreplaceability become obvious, attenuating the neuroticism.

How have the systems which host selection tournaments changed?

As has been mentioned off-handedly elsewhere in this text, while surrogation is an inescapable, eternal problem, the extent to which it suffuses a society or institution is quite consequential for that superorganism's outcomes. The present

17

age is especially suffused because it is requires, on account of a large interconnected population, high human mobility, and global coordination projects, what we'll call *distant evaluation*. Distant evaluation is in contrast with the close evaluation made possible by life in Dunbar-sized communities. Machiavelli, in *The Prince*, nodded toward this distinction with his contrasting of "sight" and "touch." The vast majority of those a prince rules will only see him at a distance; they will be subject to a very narrow peep-hole into his life and character; and the Prince's appearance in front of this peep-hole can be easily orchestrated for effect:

> ...it is unnecessary for a prince to have all the good qualities I have enumerated, but it is very necessary to appear to have them. And I shall dare to say this also, that to have them and always to observe them is injurious, and that to appear to have them is useful; to appear merciful, faithful, humane, religious, upright.

Only a few are close enough to "touch" the Prince, to live alongside him and gain some access to his less guarded character. Thus, "Every one sees what you appear to be, few really know what you are, and those few dare not oppose themselves to the opinion of the many."[18]

Within small communities, individuals are able to track reputations and debt over long periods of time. Rather than there being a single, short selection game (perhaps proceeded by some necessary preparation), one's reputation as intelligent, experienced, hard-working, etc is built in the normal process of living. This makes deception logistically and cognitively more difficult. And indeed, the outcast is archetypally distrusted by the community he newly joins—the question at

18 Machiavelli.

the top of community members' minds, of course, is what he is running from—what reputation could be disastrous enough for him to start anew. As another example, con men famously need an exit strategy (Mamet's *House of Games* provides an illustration) because maintaining long-term dissimulation (a set of fake identities, personalities, reality constructions, etc) is an exhausting proposition. Undercover spies, famously, live half-lives on account of their work.

Image is most crucial in a culture (as required by the organization of its society) of deciding from afar. Localism is protective against image manipulation precisely because it carries access to first-hand experience—rather than representations and self-representations—and because exposure (i.e. ecological monitoring) is prolonged and therefore more difficult to manipulate.[19] Thus "Over time, the facade of likability drops, and narcissists become dislikable. In a 2015 study.... impressions of the narcissists [by fellow participants] shifted from positive in the first meeting to negative rather quickly. Narcissists are built for shallow, lukewarm, and extraverted relationships."[20] In other words, we can expect them to excel at the kinds of selection games that dominate modern institutional life.

19 Robert Jackall, in Moral Mazes, somewhat relatedly notes a "managerial work ethic" which dominates American economics, where employee virtues include ability and willingness to engage in politicking, the display of hierarchical subordination to managers, and general moral flexibility (or moral subservience to company line). Jackall rather nostalgically believes these values have replaced a previous, Protestant ethic of honesty and discipline in America, and while this portrait may wax romantic, the shift from small businesses nested intimately within small communities, to large corporations nested more anonymously within global economies, doubtless changes the calculus of economic success.

20 W. Keith Campbell, *The New Science of Narcissism.*

Surrogates are not only mandated by, but help enable, social and commercial expansion. In the Upper-Middle Paleolithic Transition, human societies and economies grew increasingly complex. Trade deals and diplomacy required credible spokesmen; social hierarchies needed to be encoded in testimony for relative strangers. Fashion enters as a technology for maintaining and navigating the social graph. "By the production of symbolic artefacts that signified different social groups and kinds of relationships," David Lewis-Williams writes, "Aurignacian people were able to maintain wider networks that could exist even between people who had never set eyes on each other." The practice lends its societies an edge, spreads through the law of cultural evolution: "The surface of the body… becomes the symbolic stage upon which the drama of socialisation is enacted, and body adornment… becomes the language through which it was expressed."[21]

6.7. CODA: "SOLVING" SURROGATION

> How difficult it is not to put the sign in place of the thing; how difficult to keep the being always livingly before one and not to slay it with the word.[22]

In the end (there is no end...) surrogation is a theory of communication. The tells we communicate unwittingly; the utterances we put careful thought into. How we are read changes how we write; and when we are teased as children for some offensive bodily leakage or personal disclosure, a

21 *The Mind in the Cave: Consciousness and the Origins of Art.*

22 Goethe, *Hamburger Ausgabe*

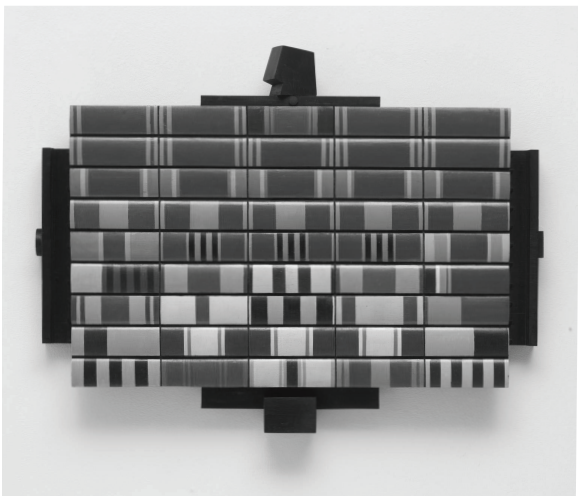
new frontier of control opens before us. When others' actions matter to us, and depend in part on our own communications—on the readings we have written—we learn to manipulate them through strategic writing, so as to better secure our welfare. We selectively suppress or expose; we imitate the writing styles of the more successful, and distance ourselves from the trappings of those who are less. We study the reactions of others, and generalize from our own internal responses, our attractions and repulsions—surrogation as theory of mind. How could we expect that corporate incentives, politics, and law would behave differently?

By extension, there is no general "solution" to surrogation, in part because surrogation is not a "problem." Rather, surrogation is a capacity, a tool, an empowering tactic which—like all powers—is limited. Instead, there are problematic approaches or attitudes toward surrogates. The reification of a single surrogate as if it "just were" the thing surrogated often leads to single-variable surrogate systems, whose thread-bareness is outperformed by more rich and comprehensive multi-variate systems.[23] Self-reporting, and surrogate evaluation by interested parties more generally, increases conflict between reporters, the reported, and the reported-to; instituting neutral third-person adjudication parties is an age-old tactic to minimize such bias.

There is also a common belief that surrogates can or should "just work"—that because their connection to the surrogateds is intrinsic or unalterable, they may be safely instituted and forgotten, staying reliable through time without

23 Amazon's leadership, for instance, tracks over 500 different performance metrics, which are then discussed, situated, and analyzed for abnormalities in Weekly Business Review (WBR) meetings. Informal evaluation systems—for instance, face-to-face human social interaction—are often far more sensitive.

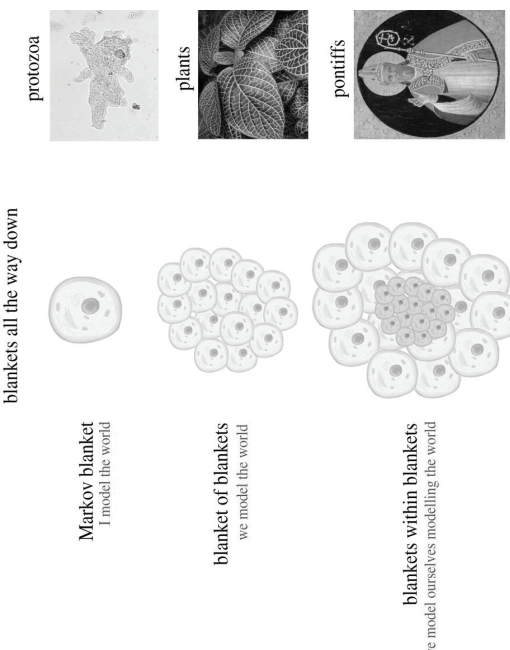
requiring oversight or critical thinking. Instead—particularly in adversarial-leaning games—the provisionality and contingency of surrogates must be constantly kept in mind. Newly adopted surrogates should be especially closely monitored, and fiercely debated by stakeholders, constantly reconciling surrogate against holistic performance (and against the performance of other surrogates). Environmental drift, and a change in the fitness of the surrogate-heuristic relative to said environment, should be seen as inevitable processes whose harms may be mitigated only through careful monitoring.



Military Person, Boris Orlov 1979. Wood relief painted in enamel. The man is replaced by the symbols of his accomplishments.

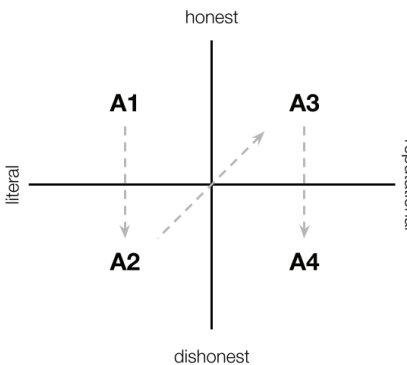
Members of a cargocult drilling with
“rifles” over their shoulders.

Surrogate guide to self-interpretation.



Part of a series of paintings commissioned for the Dominican convent of St. Catherine in Augsburg, in the late 15th C. These paintings served as surrogates for the Seven Churches of Rome, so that nuns whose health prevented in-person pilgrimage could embark on virtual ones. Participants of virtual pilgrimages (controversially at the time) could receive the same benefits of communion and indulgence as those who undertook the actual pilgrimage.

Golden calf, surrogate idol.



Temporal progressions

A1 to A2: an association between sign and consequence is formed & recognized, then exploited

A2 to A3: the slowest transition; reputational associations form between speakers and utterances, which are taken up naively

A3 to A4: reputational baggage has so accumulated that utterances cannot be taken literally even when meant literally; actors adopt whatever stance furthers goals

Temporal flow *the development of a reputation takes time*

Utterances on the left-hand side are intended or received as literal significations of their content.

Utterances on the right-hand side are intended or received through their social reputation, as signifying something more (e.g. political allegiance, covert belief) than their literal content.

Truth in comics.

Incentive	*Intended effect*	*Actual effect*
"Researchers rewarded for increased number of publications."	"Improve research productivity," provide a means of evaluating performance.	"Avalanche of" substandard, "incremental papers"; poor methods and increase in false discovery rates leading to a "natural selection of bad science" (Smaldino and Mcelreath, 2016); reduced quality of peer review
"Researchers rewarded for increased number of citations."	Reward quality work that influences others.	Extended reference lists to inflate citations; reviewers request citation of their work through peer review
"Researchers rewarded for increased grant funding."	"Ensure that research programs are funded, promote growth, generate overhead."	Increased time writing proposals and less time gathering and thinking about data. Overselling positive results and downplay of negative results.
Increase PhD student productivity	Higher school ranking and more prestige of program.	Lower standards and create oversupply of PhDs. Postdocs often required for entry-level academic positions, and PhDs hired for work MS students used to do.
Reduced teaching load for research-active faculty	Necessary to pursue additional competitive grants.	Increased demand for untenured, adjunct faculty to teach classes.
"Teachers rewarded for increased student evaluation scores."	"Improved accountability; ensure customer satisfaction."	Reduced course work, grade inflation.
"Teachers rewarded for increased student test scores."	"Improve teacher effectiveness."	"Teaching to the tests; emphasis on short-term learning."
"Departments rewarded for increasing U.S. News ranking."	"Stronger departments."	Extensive efforts to reverse engineer, game, and cheat rankings.
"Departments rewarded for increasing numbers of BS, MS, and PhD degrees granted."	"Promote efficiency; stop students from being trapped in degree programs; impress the state legislature."	"Class sizes increase; entrance requirements" decrease; reduce graduation requirements.
"Departments rewarded for increasing student credit/contact hours (SCH)."	"The university's teaching mission is fulfilled."	"SCH-maximization games are played"; duplication of classes, competition for service courses.

Advice for winning selection games typically relies on carefully controlling and deploying surrogates.

7. Appendix I: What's in a game?

Many, upon encountering a theoretical approach which frames everyday interaction as game-like, take umbrage, seeing such a frame as reductive and cold.

I can muster two broad defenses against such a critique. The first is to point to the generality of a game, as I define it. For our purposes, a game is any situation in which there is an agent with goals (preferred states) operating in an environment which furnishes obstacles and affordances. Since the goal transforms the environment, ontologically, into a landscape of obstacles and affordances; and since the fact of desire, or goal-direction, is inherent in the agent, we can simplify to say that a game is any interaction between agent and environment, i.e. it is ubiquitous. Competitive, multi-player scenarios, the abiding by provisional symbolic rules, the nesting or embedding of context "windows," and many other common game features are frequent but non-necessary criteria.

Second is to advance that we *already* implicitly view daily life as game-like, and that this is evidenced by the abundance of game-derived concepts, terms, phrases, etc that have entered common parlance. The sheer quantity of these terms is astounding; in my own research, I have been able to discover several hundred.

I have called this line of argument the lexical hypothesis—it takes as its point of departure the pragmatist notion that languge reflects use needs, and the evolutionary idea that language which is not needed or useful slowly drops out of

circulation. The abundance of game metaphors in common language—rivaled only by dramaturgical language— strikes me as strong evidence that we frequently find ourselves inside situations usefully understood through gaming lenses, for which we have requisitioned terms from sports, gambling, and warfare.

The following list is neither complete nor fastidiously checked. No doubt I have made etymological errors by including some entries. But the number of entries is, I believe, rather staggering, and gives a useful sense of overall scale:

1. 1st/2nd/3rd base
2. 3 strikes policy
3. a lot is riding on this
4. ace
5. all bets are off
6. ally
7. anybody's game
8. armchair quarterback
9. armor
10. attack
11. avant-garde
12. ball is in your court
13. ball park (figure, guess, in the)
14. bases (cover one's)
15. bat one-thousand
16. battle royale
17. battlefield (love as)
18. battleground state
19. beat back
20. beaten to the punch
21. beginner's luck
22. below par
23. below the belt (hit)
24. big league
25. bite the bullet
26. blind-sided
27. blockade
28. blow-by-blow account
29. bomb
30. boots on the ground
31. break the bank
32. bullseye
33. bunt
34. bush league
35. buzzer-beater

36. call a spade a spade
37. call it a draw
38. campaign
39. cards are stacked
40. catch flak
41. checkmate
42. choke, to
43. chomp at the bit
44. clobber
45. close call
46. clutch (performance)
47. collateral damage
48. come under fire
49. conflict (resolution)
50. count your chips
51. curveball
52. dark horse
53. deadline
54. deal (e.g. with it)
55. dealing from the bottom of the deck
56. deck stacked
57. defeat
58. defend
59. deuces
60. DLC
61. don't count me out
62. don't play games
63. double-header
64. down but not out
65. down for the count
66. down to the wire
67. DPSing
68. draw
69. drill
70. dropped ball
71. dungeon crawl
72. endgame
73. enemy
74. eye on the prize
75. face the music
76. false flag
77. fair warning
78. fences (swing for)
79. final boss
80. flop (the)
81. flop (to)
82. flying colors
83. folding
84. forced move
85. foul
86. free play
87. friendly fire
88. frontlines
89. full-court press
90. fumble
91. game (e.g. the system)

92. game (to have)
93. game day
94. game meets game
95. game plan
96. game recognizes game
97. game-changer
98. game-set-match
99. giving someone a run for their money
100. glassjaw
101. gloves off
102. go bust
103. go for broke, swing for the fences
104. god does not play dice
105. good inning
106. good sport
107. got played
108. grand slam
109. grind, to
110. ground rules
111. guessing game
112. hail mary
113. hand (weak, strong)
114. hang up your boots
115. hardball
116. have an ace up your sleeve
117. head in the game
118. heavy hitter
119. hero
120. hit me
121. hit the jackpot
122. hit-or-miss
123. hitpoint
124. hole in one
125. homefront
126. home court advantage
127. home run
128. horse race (is a)
129. hot shot
130. house rules
131. hurdles
132. hustler
133. in a league of his own
134. in the cards
135. in your wheelhouse
136. infield
137. inbounds
138. inning (top of, bottom of, ninth)
139. invasive (act, procedure, species, surgery)
140. invisible enemy
141. jeopardy
142. keep score
143. kick-off
144. knowing the deal

145. knuckle down
146. last-ditch effort
147. last man standing
148. layup
149. level playing field
150. level up
151. limit vs no-limit poker
152. loaded based
153. logistics
154. loose cannon
155. low blow
156. luck of the draw
157. magic circle
158. make the cut
159. making do with the cards you're dealt
160. marathon
161. mate
162. metagame
163. minefield
164. minigame
165. moving goalpost
166. mulligan
167. multiplayer
168. Murphy's Law
169. musical chairs
170. neck'n'neck
171. nine yards (the whole)
172. no dice
173. no holds barred
174. no man's land
175. noob
176. not the cards
177. nuclear option
178. off to the races
179. off-base
180. off-the-bat
181. offsides
182. on the block
183. on the ropes
184. only game in town
185. open vs closed world
186. opening move
187. opponent
188. orders (e.g. doctor's)
189. out of bounds
190. out of the park (knock it)
191. outfield
192. overpowered
193. own goal
194. par for course
195. pawn
196. peace (e.g. uneasy)
197. photo-finish
198. picket (line)
199. pinch hitter
200. pissing contest

201. pitstop
202. plan of attack
203. play defense
204. play dice
205. play down
206. play hardball
207. play the cards you're dealt
208. play the field
209. play the percentages
210. play the player
211. play up
212. playing for keeps
213. play your cards right
214. playoffs
215. playtest
216. poker face
217. powerup
218. pregame
219. punt
220. put all your chips in
221. put me in coach
222. put the fix in
223. PvE
224. PvP
225. quarterbacking
226. ragequit
227. rain check
228. raise someone
229. raise the stakes
230. rally
231. rat race
232. recon
233. referee
234. reinforce/ments
235. respawn
236. retreat (e.g. tactical)
237. ride someone's coattails
238. rival
239. roll of the dice
240. rolling with punches
241. rope-a-dope
242. ropes (learn the, know the)
243. royal flush
244. rules lawyering
245. run out the clock
246. run the table
247. running interference
248. running point
249. salvo (opening, closing)
250. save point
251. saved by the bell
252. score (v.)
253. see the whole board
254. seventh inning stretch
255. shell-shocked
256. shot (take your)

257. showing your hand
258. shuffling
259. sidequest
260. sin (archery)
261. single-player
262. sitting on the bench
263. six (watch your)
264. skin in the game
265. slam dunk
266. snipe
267. softball
268. sport (bad, good)
269. sportsmanship
270. sprint
271. stalemate
272. stay the course
273. stepping up to the plate
274. sticky wicket
275. strategy (a winning, a losing)
276. strike out
277. sucker punch
278. suit up
279. sweeten the pot
280. tactic
281. target
282. take a mulligan
283. take one for the team
284. take your shot
285. team player
286. the economy's a casino
287. throw the game
288. tilted
289. tip your hand
290. touch base
291. touchdown
292. trench warfare
293. troops (rally the)
294. truce
295. two can play that game
296. under the wire
297. underpowered
298. up the ante
299. victor/y
300. wallop
301. war (price, on cancer, all's fair in love and)
302. warning shot
303. whale
304. wheelhouse (in one's)
305. when the chips are down
306. whistleblower
307. white flag (wave a)
308. wildcard
309. winner's curse
310. winning hands down

8. Appendix II: Material Concepts

Here I provide a brief overview of concepts related to surrogation, in the hope that they might help facilitate future theoretical and synthetic work around the critical role of representation in incentive systems and games of strategy.

Wireheading is a speculative problem in the design of artificial intelligence whereby an AI discovers ways to hack its reward function such that rewards are dispensed without the AI needing to accomplish the work which the reward function was designed to incentivize. The *underspecification problem* refers to the difficulty (or impossibility) of fully specifying every situation, and desired behavior, from an artificial intelligence; it bears similarities to my discussion of letter and spirit. A *nearest unblocked strategy* is the idea that an AI, if blocked from pursuing some desired course of action, will ruthlessly search for the nearest unblocked (technically allowed) course which most closely accomplishes its desires.

Campbell's Law is the idea 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 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." The superset of Campbell's Law is *Goodhart's Law*, most frequently paraphrased as, "When a measure becomes a target, it ceases to be a good measure." For instance, measurements of crime,

when optimized toward by police, lose validity and insight as measurements.

Robert K. Merton, in 1940's "Bureacratic Structure and Personality," introduced the concept of *goal displacement*, by which "an instrumental value becomes a terminal value." As metrics are established to evaluate performance, they displace the original institutional goals and become the ends optimized for by embedded agents. Merton's frame precedes Goodhart's Law by several decades while making an almost identical observation: that "when certain indicators get officially, or quasi-officially, established as measures of this, that, or the other, there will be, one should look for, efforts to manipulate the numbers by one's behavior." He scoped the problem specifically to academic citations, prophesying that as "more and more citations are used both officially and unofficially as measures of contribution" and "relative standing," citation behavior among academics will actively change. This manipulation leads the indicator to "no longer indicate what it once did."

Nguyen's theory of *value capture*, as outlined in 2020's *Games: Agency As Art*, recapitulates Merton and Goodhart. Nguyen defines value capture as the substitution of a simplified metric or indicator for a richer holistic value—for instance, the concrete and objective "counting steps" for the vague but holistic "staying in shape." Nguyen distinguishes it from Goodhart's Law in that the substitution is internalized by the agents situated within the surrogate incentive structure. Venkatesh Rao's *gollumization*, a reference to Tolkien's Gollum character, similarly describes a "hollowing out" of a person's holistic value structure by single-minded or "fetishistic" addiction.

The *cobra effect* refers to an apocryphal example of perverse incentives in which a colonial city (often Delhi) is said to have offered a bounty on slain cobras to combat an urban infestation. This, of course, leads locals to breed cobras for resale to the colonial government, exacerbating the problem.

In ethology, a *signal* is information which evolution has selected an animal to emit, because its production alters the behavior of other animals in a way advantageous to its own reproduction or survival. A *cue* is information produced by an animal which is not advantageous to it (e.g. a mouse rustling grass as it passes through a field), produced as a byproduct of other advantageous actions, which is used by observing organisms to inform their own respective actions (e.g. a hunting owl). *Mimicry* is the free-riding of an honest signal, such as bright red coloration to signal toxicity, without possessing the underlying traits signaled. It is now established wisdom,in ethology that a complete absence of mimicry is not a stable equilibrium—in other words, that some amount of mimicry is inevitable over the longue durée. This echoes economist Dan Davies's observation that some amount of fraudulence, in an economic system, is not only inevitable but also economically desirable. (There is a point of diminishing returns at which the cost, to a system, of stamping out fraud exceeds the cost of the fraud itself.)

In economics, Joseph Stiglitz and Andrew Weiss have theorized a set of games in which an informed player (i.e. one who knows the value of the trade) interacts with an uninformed player (i.e. one who doesn't) in attempting to secure an offer (i.e. a purchase price). A *signaling game* refers to such an interaction where the informed player moves first, signaling the value of his offered good; a *screening game* involves the

uninformed player moving first. These terms are indebted to the similar game-theoretic concepts.

Elsewhere in the field of economics, *search theory* and *matching theory* refer to the study of buyers or sellers in their ongoing attempts to find trading partners, and their use of *attributes* or signals to judge potential partners. Job hiring, bank loans, and traditional markets are typical domains theorized.

A *proxy* is used in statistical analyses to measure some underlying, but unobservable, overly abstract, or difficult to quantitize phenomenon. This may be referred to as the *operationalization* of the underlying phenomenon (or *latent variable*) via an *observable* or *manifest variable*. In psychometrics, *construct validity* or *test validity* refers to the extent to which a proxy can be relied on as a reflection of some underlying phenomenon. In information theory, *mutual information* reflects the extent to which two variables are mutually dependent, that is, to which information from one known variable can be used as the basis of inference about the other, unknown variable. Similarly, statistics treats *Fisher information* as describing the extent to which an observable random variable can predict an unknown parameter. No doubt a better treatment of—or paradigm for—the surrogation idea would more dramatically integrate these statistical and information-theoretic concepts.

9. Bibliography

Alexander, Scott. "Book Review: Fussell On Class," *Astral Codex Ten* 2021.

Apstein and Prewitt. "The New Steroids," *Sports Illustrated* 2021.

Arbital. "Nearest Unblocked Strategy," *Arbital.*

Assumption, Feast of. "PvP v PvE," *TIS* 2022.

Banana, Literal. "Ignorance, a Skilled Practice," *Carcinisation* 2020.

Bateson, Gregory. *Steps To An Ecology of Mind* 1972.

Bernard, Jessie. "The Theory of Games of Strategy as a Modern Sociology of Conflict," *American Journal of Sociology* 1954.

Bhagwat, Sam. "Playing Games to Leave Games," *Ribbonfarm* 2014.

Bishop, Michael. "The Possibility of Conceptual Clarity in Philosophy," 1992.

Blogospheroid. "The Importance of Goodhart's Law," *AI Alignment Forum* 2010.

Boluk and LeMieux. *Metagaming: Playing, Competing, Spectating, Cheating, Trading, Making, and Breaking Videogames* 2017.

Buckner, William. "Charlatinism: Realms of Deception and Religious Theater," *Traditions of Conflict* 2019.

Carse, James. *Finite & Infinite Games* 1986.

Chapman, David. *Meaningness*.

Chicken, Crispy. "Wireheading Is a Teleological Misnomer," Substack 2021.

Chin, Cedric. "Goodhart's Law Isn't as Useful as You Might Think," *Commoncog* 2023.

Choi, Jongwoon, Hecht & Tayler. "Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation," *The Accounting Review* 2012.

———. "Strategy Selection, Surrogation, and Strategic Performance Measurement Systems," *Journal of Accounting Research* 2013.

Christian and Griffiths. *Algorithms to Live By* 2016.

Concierge, Hotel. "How To Be Attractive," *Hotel Concierge* 2016.

———. "The Tower," *Hotel Concierge* 2017.

Connable, Ben. *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency* 2012.

Constantin, Sarah. *Otium*.

Danto, Arthur. *Warhol* 1997.

Davies, Dan. *Lying for Money* 2022.

deBoer, Frederick. Substack.

de Falco, Gianni. "Incentives & degenerate play," *TIS* 2022.

———. "Not all heroes wear capes," *TIS* 2022.

DeDeo, Simon. "Information Theory for Intelligent People" 2018.

Elwood, Zachary. "Jury selection strategies, with Christina Marinakis," *People Who Read People: A Behavior and Psychology Podcast* 2018.

EmpLemon. *there will Never Ever be another Melee player like Hungrybox*, YouTube 2020.

Espeland & Sauder. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability* 2016.

Evangelista, Nick. *The Inner Game of Fencing* 2000.

Falkenstein, Eric. "A Batesian Mimicry Explanation of Business Cycles," *Falkenblog* 2010.

Farley, Patrick D. *A World of Symbols* 2020.

Feynman, Richard. "Caltech Commencement Address" 1974.

Fitzgerald, Neil. "Board games are a social construct," *TIS* 2022.

Fleischer, Matthew. "Dangerous Jails," *WitnessLA* 2011.

Focke, Kevin. *Essentialized* 2020.

Forrester, John. *Truth Games: Lies, Money, and Psychoanalysis*. 1997.

Friston et al. "The Markov Blankets of Life," *Journal of The Royal Society Interface* 2018.

Fussell, Sam. *Muscle* 1991.

Gioia and Corley. "Being Good vs. Looking Good: Business School Rankings and the Circean Transformation from Substance to Image," *Academy of Management Learning and Education* 2002.

Goffman, Erving. *Strategic Interaction* 1969.

———. *The Presentation of Self in Everyday Life* 1956.

Greer, Nick. "Private Inequity," *Lesser Works* 2023.

Guzey, Alexey. "Reviving Patronage and Revolutionary Industrial Research" 2019.

Hanson and Simler. *The Elephant in the Brain* 2017.

Hazard, Natural. "Arguing Definitions As Arguing Decisions," theinexactsciences.github.io 2021.

———. "Gödel's Legacy: A game without end," *LessWrong* 2020.

———. "Intuitionistic Type (of Guy) Theory" 2021.

Hayakawa, S. I. *Language in Thought and Action* 1949.

Hubinger et al. "Risks from Learned Optimization in Advanced Machine Learning Systems," preprint 2019.

Huemer, Michael. "In Praise of Passivity," *Studia Humana* 2012.

Inbar and Inzlicht. *Two Psychologists Four Beers* 2018-2021.

Inzlicht, Michael. "Reckoning with the Past" 2016.

Lysford, Collin. *Desystemize*.

Jackall, Robert. *Moral Mazes: The World of Corporate Managers* 1988.

Jameson, A D. "Experimental Fiction as Genre and as Principle," *Big Other* 2010.

———. "Punk Ethos & The Blog." Interview, *Suspended Reason* 2018.

Kahneman and Tversky. "On the Psychology of Prediction." *Psychological Review* 1973.

Kilcullen, David. *Counterinsurgency*. 2010.

Kuran, Timur. *Private Truths, Public Lies* 1997.

Lantz, Frank. "Donkeyspace," *Game Design Advance* 2011.

——— and Zimmerman. "Rules, Play and Culture: Towards an Aesthetic of Games" 2012.

Latterner, Tim. "Paul Sevigny Won't Stop Partying," *GQ* 2021.

Lewis-Williams, David. *The Mind in the Cave: Consciousness and the Origins of Art* 2004.

Lyotard, Jean-François. *The Postmodern Condition: A Report on Knowledge* 1984.

Luu, Dan. "Individuals matter" 2021.

Keith Campbell, W. *The New Science of Narcissism* 2020.

Machiavelli, Niccolò. *The Prince* 1532.

Mackay, Robin. "Hyperplastic-Supernormal" 2016.

Manheim, David. "Goodhart's Law and Why Measurement Is Hard," *Ribbonfarm* 2016.

———. "Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence," *Big Data and Cognitive Computing* 2019.

———. "Overpowered Metrics Eat Underspecified Goals," *Ribbonfarm* 2016.

Massey, Mooney, Torres, and Charles. "Black Immigrants and Black Natives Attending Selective Colleges and Universities in the United States," *American Journal of Education* 2007.

Merton, Robert. "Bureacratic Structure and Personality" 1940.

Modernist, Possible. "Degenerate play," *TIS* 2022.

———. "Generalized hacking," *TIS* 2022.

———. "Kingmakers," *TIS* 2022.

Muller, Jerry. *The Tyranny of Metrics* 2018.

Nguyen, C. Thi. "Games and the Art of Agency," *The Philosophical Review* 2019.

———. *Games: Agency As Art.* 2020.

Nietzsche, Friedrich. *On Truth and Lies in a Nonmoral Sense* 1873.

Nightjack. "Goalposts Moving" 2008.

Noë, Alva. *Strange Tools: Art and Human Nature* 2015.

OpenAI. "Faulty Reward Functions in the Wild," *OpenAI Blog* 2016.

Ortega and Maini et al. "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," *DeepMind Safety Research* 2018.

Perry, Sarah. "Meaning and Pointing," *Ribbonfarm* 2015.

———. "Puzzle Theory," *Ribbonfarm* 2015.

Pizarro and Sommers. *Very Bad Wizards* 2012-2021.

Rao, Venkatesh. *The Gervais Principle* 2009.

———. *The Gollum Effect* 2011.

Reason, Suspended. "Conceptual Engineering: The Revolution in Philosophy You've Never Heard Of," *LessWrong* 2020.

———. "Intro to Cargocult." *Suspended Reason* 2016.

———. "Letter to Tamler Sommers and David Pizarro," *Suspended Reason* 2021.

———. "On generalized reading." *LetterWiki* 2021.

———. "The Dark Miracle of Optics," *Suspended Reason* 2020.

———. "The Tyranny of Round Numbers," *Carcinisation* 2017.

Rimer and Arenson. "Top Colleges Take More Blacks, but Which Ones?" *The New York Times* June 24 2004.

Rodamar, Jeffery. "There Ought to Be a Law! Campbell versus Goodhart," *Significance* 2018.

Salen Tekinbaş and Zimmerman. *Rules of Play: Game Design Fundamentals* 2003.

Schegloff, Emanuel A. "Confirming Allusions: Toward an Empirical Account of Action," *American Journal of Sociology* 1996.

Schelling, Thomas C. *The Strategy of Conflict* 1960.

Sciutto, Cristóbal. "Lacunae, DIY, and gestalts" 2021.

Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* 2008.

Shah, Rohin. "Mesa Optimization: What It Is, and Why We Should Care." *LessWrong* 2019.

Simler, Kevin. "Minimum Viable Superorganism." *Ribbonfarm* 2016.

Sirlin, David. *Playing To Win* 2005.

Simon & Burns. *The Wire* 2002-2008.

Strauss, David. "The Anti-Formalist," *University of Chicago Law Review* 2007.

Sutherland, Rory. Interview with R. Roberts, *EconTalk* 2019.

Taleb, Nicholas Nassim. *Incerto* 2001-2018.

Taylor, Tim. *Conversable Economist.*

Traldi, Oliver. "Them Among Us," *Arc Digital* 2020.

TVTropes. "Rules Lawyer." Accessed 2023.

U.S. v. Marshall, 908 F.2d 1312 (7th Cir. 1990).

Vollmer, Hendrik. "What Kind of Game Is Everyday Interaction?" *Rationality and Society* 2013.

Weick, Karl E. *The Social Psychology of Organizing* 1969.

Wilson, James Q. *Bureaucracy* 1989.

Wittgenstein, Ludwig. *Philosophical Investigations* 1953.

Yarkoni, Tal. "The Generalizability Crisis" 2019.

10. INDEX (OUT OF DATE)