Behavioral and Brain Sciences

cambridge.org/bbs

Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems

Yohan J. John[a], Leigh Caldwell[b], Dakota E. McCoy[c,d,e] and Oliver Braganza[f,g]

[a]Neural Systems Laboratory, Department of Health and Rehabilitation Sciences, Boston University, Boston, MA, USA; [b]Irrational Agency, London, UK; [c]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA; [d]Hopkins Marine Station, Stanford University, Pacific Grove, CA, USA; [e]Department of Biology, Duke University, Durham, NC, USA; [f]Institute for Experimental Epileptology and Cognition Research, University of Bonn, Bonn, Germany and [g]Institute for Socioeconomics, University of Duisburg-Essen, Duisburg, Germany
yohan@bu.edu
leigh@irrationalagency.com
mccoy6@stanford.edu

Target Article

Cite this article: John YJ, Caldwell L, McCoy DE, Braganza O. (2024) Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences* **47**, e67: 1–56. doi:10.1017/S0140525X23002753

Target Article Accepted: 8 May 2023
Target Article Manuscript Online: 26 June 2023
Commentaries Accepted: 13 October 2023

Keywords:
Goodhart's law; Campbell's law; goal displacement; the cobra effect; the McNamara fallacy; addiction; signalling; the proxy treadmill; proxyeconomics

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 20) and an Authors' Response (p. 47). See bbsonline.org for more information.

Corresponding author:
Oliver Braganza;
Email: oliver.braganza@ukbonn.de

Abstract

When a measure becomes a target, it ceases to be a good measure. For example, when standardized test scores in education become targets, teachers may start "teaching to the test," leading to breakdown of the relationship between the measure – test performance – and the underlying goal – quality education. Similar phenomena have been named and described across a broad range of contexts, such as economics, academia, machine learning, and ecology. Yet it remains unclear whether these phenomena bear only superficial similarities, or if they derive from some fundamental unifying mechanism. Here, we propose such a unifying mechanism, which we label *proxy failure*. We first review illustrative examples and their labels, such as the "cobra effect," "Goodhart's law," and "Campbell's law." Second, we identify central prerequisites and constraints of proxy failure, noting that it is often only a partial failure or *divergence*. We argue that whenever incentivization or selection is based on an imperfect proxy measure of the underlying goal, a pressure arises that tends to make the proxy a worse approximation of the goal. Third, we develop this perspective for three concrete contexts, namely neuroscience, economics, and ecology, highlighting similarities and differences. Fourth, we outline consequences of proxy failure, suggesting it is key to understanding the structure and evolution of goal-oriented systems. Our account draws on a broad range of disciplines, but we can only scratch the surface within each. We thus hope the present account elicits a collaborative enterprise, entailing both critical discussion as well as extensions in contexts we have missed.

1. Introduction

In the spring of 1902, French colonial officials in Hanoi, fearful of the bubonic plague, declared war against a self-inflicted rat infestation (Vann, 2003). Officials incentivized rat catchers by offering a reward for every delivered corpse. In the following months the number of delivered rat corpses increased exponentially, yet the underground population seemed unaffected. As the heaps of corpses grew and became a nuisance, officials started rewarding the delivery of rat tails rather than whole animals. At the same time the city expanded its incentive scheme to the general population, promising a 1 cent bounty for every tail delivered. Residents quickly began to deliver thousands of tails. However, increasing numbers of rats without tails were soon observed scurrying through the city, perhaps left alive to breed and therefore supply the newly valuable tails. Even worse, enterprising individuals began to breed rats, farming the tails in order to collect rewards more efficiently.

This is an example of what is known in economics as Goodhart's law: "When a measure becomes a target, it ceases to be a good measure" (Goodhart, 1975; Strathern, 1997). The measure (rat corpses or tails) is an operational *proxy* for some *goal* or purpose (reduction of the rat population). However, when it becomes the target, its correlation with this goal decreases, or in extreme cases, disappears, leading to unintended and often adverse outcomes. In this case the rat population of Hanoi ended up surging when the programme was terminated: The now-worthless rats were set free in the city. The proxy failed both to approximate and to guide goal-oriented action. As we will demonstrate, "Goodhart-like" phenomena have been discovered and rediscovered

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

across a broad range of contexts and scales, ranging from centralized governance to distributed social systems and from evolutionary competition to artificial intelligence (see Table 1). Although the physical mechanisms vary from case to case, there are several structural features that recur throughout them (see, e.g., Biagioli & Lippman, 2020; Braganza, 2022b; McCoy & Haig, 2020), indicating that the similarities are not superficial.

Here, we will draw out these structural commonalities to reveal a single core phenomenon: When the pursuit of a goal depends on the optimization of an imperfect proxy in a complex system, a pressure emerges that pushes the proxy away from the goal. *Optimization* here means that some regulatory feedback mechanism promotes the increase of the proxy via incentivization, reinforcement, selection, or some other means. We propose that the disparate "Goodhart-like" phenomena can be unified under the more descriptive term *proxy failure*. Similar to the notion of "market failure," the term is not meant to imply a *complete* failure, but only a *divergence* of the proxy from the underlying goal. The aim of this paper is twofold: (1) to set up a unified theoretical perspective that facilitates recognition of proxy failure across biological and social scales, and (2) to explore implications of this perspective. In particular, recognizing how the mechanisms and constraints of proxy failure between diverse systems overlap or differ should help to understand when and how it can be mitigated, or indeed harnessed.

The remainder of the paper is structured as follows. Section 2 briefly reviews examples of proxy divergence across contexts. Section 3 extracts and analyses the common pattern underlying these examples: Proxy-based optimization of regulatory goals. Section 4 elaborates these ideas in three concrete contexts: neuroscience (where we are, to the best of our knowledge, the first to apply the framework of proxy failure), economics, and ecology. Section 5 discusses consequences of proxy failure across biological and social systems, including how it can drive complexity and how it affects the structure and behaviour of nested hierarchical systems. Section 6 concludes.

Yohan J. John research assistant professor in the Neural Systems Laboratory, Boston University, is a computational neuroscientist who develops models of the neural circuits underlying cognitive–emotional interaction and psychiatric disorders. His work has been funded by grants from the US National Institutes of Health.

Leigh Caldwell director of the research consultancy Irrational Agency, works on cognitive economics across both commercial and scientific applications. He is author of *The Psychology of Price* (Crimson Publishing, 2012), certified member of the Market Research Society, and has authored publications on narrative modelling, quantitative modelling of choices, and behavioural macroeconomics.

Dakota E. McCoy is a Stanford Science Fellow who has authored papers across the environmental, evolutionary, and physical sciences. McCoy research includes subtle arms races in human pregnancy, coral reefs, and more. She has received the Dobzhansky Prize for contributions to evolutionary research, the Early Career Investigator Award from the American Society of Naturalists, and funding from NSF and DOD. McCoy is a Rhodes Scholar.

Oliver Braganza a postdoctoral researcher at the University of Bonn (Institute for Experimental Epileptology and Cognition Research) and Duisburg-Essen (Institute for Socioeconomics), has authored papers across experimental and theoretical neuroscience and meta-science/economics. His transdisciplinary work on "proxyeconomics" was kickstarted by an innovation grant by VW-foundation (Originalitätsverdacht). His neuroscientific work has been funded by grants from BonFor and the DFG.

2. A brief overview

The best-known label for proxy failure may be Goodhart's law, traced to the economist Charles Goodhart in the context of monetary policy (Goodhart, 1975). However, the law is known to be closely related to a jumble of other laws, fallacies, or principles (Table 1), some of which predate Goodhart (Csiszar, 2020; Merton, 1940; Rodamar, 2018). They occur when a regulator or administrator with some goal incentivizes citizens or employees. The latter then identify weaknesses that allow them to "game" or "hack" the incentive system *intentionally* (Table 1, yellow area). For instance, the McNamara fallacy, the Lucas critique, and Goodhart's law all involve governmental incentive schemes that are "gamed" by soldiers or citizens. We write "gamed" in scare-quotes, because agents can also be understood to simply be rationally responding to incentives (Lucas, 1976). The economics literature abounds with analogous cases: A "principal's" incentive schemes are "gamed" by rational agents (Baker, 2002; Bénabou & Tirole, 2016; Bonner & Sprinkle, 2002; Kerr, 1975; we will discuss this in more detail in sect. 4.2).

In such social systems, it is clear that human intentions play a key role in proxy failure. But recent research in machine learning and artificial intelligence (AI) has shown that intentional gaming is not necessary for proxy failure to occur. Instead, it is a consequence of proxy-based optimization in itself (Table 1, blue area; Amodei et al., 2016; Ashton, 2020; Beale, Battey, Davison, & MacKay, 2020; Manheim & Garrabrant, 2018). The AI literature is beginning to uncover the causal (Ashton, 2020; Everitt, Hutter, Kumar, & Krakovna, 2021) and statistical (Beale et al., 2020; Zhuang & Hadfield-Menell, 2021) foundations of the phenomenon. Indeed, AI appears to be an ideal microcosm to study Goodhart's law, its variants (Demski & Garrabrant, 2020; Manheim & Garrabrant, 2018), mitigation strategies (Amodei et al., 2016; Everitt et al., 2021; Thomas & Uminsky, 2022), and the potentially severe adverse social consequences (Bostrom, 2014; O'Neil, 2016; Pueyo, 2018; Zuboff, 2019). A full analysis of proxy failure in AI is beyond the scope of the present paper; see the citations above for excellent treatments.

An additional set of studies raises the question of whether a well-defined human regulator with a clear goal is required for proxy failure to occur. These studies describe proxy failure in distributed social systems, such as academia or markets, where the goal may be a dynamically changing social agreement (Table 1, red area; Biagioli & Lippman, 2020; Braganza, 2022b; Fire & Guestrin, 2019; Poku, 2016). For instance, the academic publication process (Biagioli & Lippman, 2020; Braganza, 2020; Smaldino & McElreath, 2016), education (Nichols & Berliner, 2005), healthcare (O'Mahony, 2018; Poku, 2016), and competitive markets (Braganza, 2022a, 2022b; Mirowski & Nik-Khah, 2017) appear to be subject to proxy failure, even though it is difficult to clearly define the respective goals. To the best of our knowledge, the phenomenon has not yet been explicitly explored in democratic elections, but practices such as "vote buying" (Finan & Schechter, 2012) or insincere election promises (Thomson

Table 1. Examples of proxy failure

	Example/name	Goal	Proxy	Agents	Regulator	Failure claim
Governance	Monetary policy **Goodhart's law**	Economic regulation	Financial assets	Traders/banks	Government	Goodhart (1975)
	Education **Campbell's law**	Knowledge, skills	Standardized test scores, grades	Teachers, schools	Government, funders	Campbell (1979); Koretz (2008); Nichols and Berliner (2005); Stroebe (2016)
	Macroeconomics **Lucas critique**	Economic growth	Interest-, inflation rate	Market participants	Government	Lucas (1976)
	Military **McNamara fallacy**	War victory (Vietnam War)	Body count	Soldiers	Government-, military leadership	Yankelovich (1972)
	Cobra effect	Fewer cobras	Dead cobras	Citizens	Government	Siebert (2001)
	Management **Indicatorism**	Profit, firm value	KPI, quarterly returns, etc.	Employees, subdivisions	Corporation, manager	Baker (2002); Kerr (1975); van der Kolk (2022)
	Bureaucracy **Goal displacement**	Arbitrary original goal	"Instrumental value"	Lower-level bureaucrats	Higher-level bureaucrat	Griesemer (2020); Merton (1940); Muller (2018)
AI	**Unethical optimization**	Success in an ethical way	Objective function	Potential strategies	AI architecture	Beale et al. (2020)
	Reward tampering	Arbitrary AI goal	Objective function	Potential outcomes	AI architecture	Everitt et al. (2021); Manheim and Garrabrant (2018)
	Social media	E.g., entertainment	No. of clicks, time on platform	Content (e.g., videos)	Social media corporation	Bessi et al. (2016); Faddoul, Chaslot, and Farid (2020)
	Search engine optimization	Search relevance	Search algorithm	Websites	Search engine provider	Bradshaw (2019); Ledford (2016)
Society	Science	Quality research	Publication metric	Researchers, labs	Funders, universities	Biagioli and Lippman (2020); Braganza (2020)
	Economics	Prosperity, wellbeing	Profit, GDP	Companies	Market	Braganza (2022a, 2022b); Kelly and Snower (2021)
	Politics	Good governance	Votes, popularity	Parties, politicians	Election	Finan and Schechter (2012); Thomson et al. (2017)
	Medicine	Quality healthcare	Patient numbers, profit	Doctors, hospitals	Market, funders	O'Mahony (2018); Poku (2016)
Ecology	Embryo selection (primates and horses)	Offspring quality	Chemical signal	Embryo	Parent	McCoy and Haig (2020)
	Embryo selection (plants)	Offspring quality	Chemical signal	Embryo	Parent	Shaanker et al. (1988); Willson and Burley (1983)
	Sexual selection	Mate fitness	Sexual signal	Displaying sex	Choosing sex	Albo, Winther, Tuni, Toft, and Bilde, (2011); Backwell, Christy, Telford, Jennions, and Passmore (2000); Funk and Tallamy (2000); Gasparini et al. (2013)
	Runaway niche construction	Biological, cultural fitness	Physical, behavioural trait	Selected trait	Constructed niche	Rendell, Fogarty, and Laland (2011)
	Neonate selection (marsupials)	Offspring quality	Speed to find teat	Neonate	Mother	Present paper

(*Continued*)

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Table 1. (*Continued.*)

	Example/name	Goal	Proxy	Agents	Regulator	Failure claim
Neuroscience	Preference learning	Utility, fitness	Reward signal (e.g., dopamine)	Preferences, habits	Organism, meta-cognition	Present paper
	Diet	Nutrition, health	Sweetness, saltiness reward	Food representations		
	Addiction	Learning	Dopamine bursts	Plan-, habit representations		
	Exploration	Knowledge	Novelty-related reward signal	Plan representations		

Examples from different literatures and across scales, in which Goodhart's law or one of its analogues has been explicitly or implicitly invoked. Note that many examples refer to reviews or theoretical integrations of primary literature within the respective fields. Instances, which have led to the coining of "laws" (or similar), have been highlighted in boldface. Rows with no boldface entry generally refer to at least one such "law". Our examples can be categorized into five domains highlighted by colour. Yellow: social systems with central regulator; blue: machine learning or artificial intelligence research; red: distributed social systems relying on metrics; green**:** ecological systems; grey: neuroscience.

et al., 2017) clearly fit the pattern. Building on these insights, Goodhart's law has recently been invoked in yet another context, namely ecology (Table 1, green area; McCoy & Haig, 2020). In the present paper, we add neuroscience to the list (Table 1, grey area; sect. 4.1).

Together, these examples raise the question of whether proxy failure is a fundamental risk in any complex, goal-oriented system.

3. A unified theoretical perspective

Despite the striking similarity between the diverse phenomena in Table 1, a unified explanation has not yet been offered. To this end, we first set up a consistent terminology (sect. 3.1). Briefly, we frame proxy failure in terms of a *regulator* with a *goal*, who uses a *proxy* to incentivize or select *agents* in service of this goal. Our account can be summarized in the following propositions and corollaries (Table 2), which we expand on in the remainder of the paper.

3.1. A unifying terminology: Regulator, goal, agent, proxy

To facilitate comparisons across disparate instances of proxy failure, we define four inter-related terms: *regulator*, *goal*, *agent*, and *proxy* (Box 1). A *regulator* is any entity or system with an apparent *goal*. To pursue this goal, the regulator incentivizes or selects *agents* based on some *proxy*. The agents in turn pursue the proxy, or are selected on the basis of the proxy. We will now examine each term in sequence.

A *regulator* is the part of the system that pursues a goal by influencing the agents' behaviour or properties using some form of feedback. The regulator predicates feedback on agents' performance in terms of the proxy. As we will explore below, the regulator may not only be a distributed system (e.g., a market) or an institution (a government), but it can also be a person, a neural subsystem (the dopamine system), or an ecological subpopulation (peahens). Conscious awareness is not required for regulatory feedback, which can occur via selection or reinforcement (rewards and punishments). In the case of the Hanoi rat massacre, the regulator was the French colonial government, which pursued its goal of reducing the rat population through feedback in the form of monetary incentives. In the context of sexual selection, the regulator is the population of potential mating partners and the feedback is access to mating opportunities.

A *goal* refers to the state of the system that the regulator seeks to establish. In human social contexts, such goals are sometimes

Table 2. Key propositions and their corollaries

Propositions	
A	A regulatory system with a goal will typically use a proxy (or set of proxies) to pursue it (sect. 3.2).
B	The use of the proxy for regulatory feedback creates a pressure towards proxy failure, i.e., a pressure towards divergence between proxy and goal (sect. 3.3).
C	The proxy diverges from the goal when this pressure is not counterbalanced or constrained by the system (sect. 3.3).
Corollaries	
1	Proxies and proxy failure are ubiquitous in both artificial and natural systems and occur across scales ranging from the social to the intrapersonal (sects. 4.1–4.3).
2	The language typically used to describe proxy failure often implies human intentions; consider "hacking" or "gaming." But proxy failure can be generalized as a mechanistic process without requiring intentions or awareness (sects. 4.1–4.3).
3	The interaction of pressures towards and against proxy failure has predictable consequences across systems and scales. These include *instability*, *inflation*, and *elaboration* within complex hierarchical organizations (sect. 5.1).
4	Considering proxies and proxy failure across levels of nested hierarchical systems suggests novel perspectives to the organization of living systems. For instance, regulation can function via a cascade of proxies through the levels of a nested hierarchy (sect. 5.2).
5	The continued existence of apparent proxy failure in a system may indicate slack in countervailing pressures. However, sometimes proxy failure at a lower level can be coopted by a higher-level entity to serve the higher-level goal (sect. 5.3).

Box 1. Glossary of terms

A *regulator* is any entity with an apparent *goal*. To pursue this goal, the regulator incentivizes or selects *agents* based on some *proxy*.

Regulator: the part of the system that pursues a goal by influencing the agents' behaviour or properties using some form of feedback.

Goal: the state of the system that the regulator seeks to establish.

Agent: an entity, process, or abstract system that is the target of regulation.

Proxy: an output or property of each agent that the regulator uses to approximate the goal. This may be a *cue or signal* in biology, or a *performance metric or indicator* in social contexts.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

explicit and simple (reducing a rat population, increasing shareholder value) and sometimes difficult to precisely define (social welfare, improved education, or scientific progress). The degree to which goals are captured by their respective proxy measures may vary considerably (Braganza, 2022b), even though many proxies were formulated explicitly to quantify abstract goals (e.g., test scores in education or the journal impact factor in academia). In nonhuman contexts, such as in ecology, explicit goals cannot generally be assumed. A peahen may act *as if* to optimize her offspring's fitness, but she might not consciously represent or monitor this goal. In such cases, imputing goals nevertheless remain an efficient way to account for empirical findings, a perspective known as teleonomic reasoning (Mayr, 1961; Appendix 1).

An *agent* is an entity, process, or abstract system that is the target of regulation. Agents are subject to competition, selection, or reinforcement on the basis of some output or property. They need not be human or conscious (or even active in any sense), but they must possess multiple ways of producing or expressing a proxy, which can be influenced by feedback from the regulator. In the case of the Hanoi rat massacre, the agents were the people bringing in rat corpses and tails. In educational systems, the agents are teachers who end up "teaching to the test." In the brain, agents are neural representations "competing" for neural resources (see sect. 4.1). In sexual selection, agents are peacocks trying to appear attractive to choosy peahens.

A *proxy* is an output or property of each agent that the regulator uses as an approximation of the goal, or the extent to which a goal is achieved. An agent's "performance" in terms of the proxy is the key factor determining the regulator's feedback. Examples of proxies (and their respective goals) include: The number of rat tails (approximating rat population size) in the Hanoi example; test-scores (approximating educational quality); peacock plumage (approximating offspring quality); and dopamine signals (approximating value in decision making). The proxy can typically be expressed as a scalar metric (i.e., a magnitude). This scalar may be aggregated from multidimensional inputs or serve as a summary of multiple metrics. The regulator uses the proxy to adjust its feedback to the agents (reward, ranking, selection, or breeding opportunity). This induces agents to increase proxy production, either actively or via passive selection. The term *proxy* focuses attention on the fact that the metric should *approximate* a goal, and that the fidelity of approximation is key.

A regulator with respect to one system may be an agent when viewed as part of a different system. For instance, in many cases of sexual competition, males are regulated by female choices, but females are themselves competing against each other. Furthermore, agents and regulators may be organized in nested hierarchies (Ellis & Kopel, 2019; Haig, 2020; Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018), where agents at one level are regulators to a lower level (consider a corporation where general management regulates departments, which in turn regulate subdepartments, etc.; more on this in sect. 5). In this context we should note that we will employ anthropomorphic language in the description of both human and nonhuman systems, simply because it is useful and concise (Ågren & Patten, 2022; Dennett, 2009; Appendix 1).

3.2. Why proxies are imperfect: Limits of legibility and prediction, and the necessity to choose

It is natural to wonder if a regulator could dispense with proxies and focus directly on the goal. We argue that this is often very difficult or impossible. Proxies are almost inevitable in complex goal-oriented systems. A regulator, even if it takes the form of a distributed process, faces three limitations: Restricted legibility; imperfect prediction; and the necessity to choose. These limitations also apply to agents.

First, regulators are limited by *legibility* (Scott, 2008). Many abstract goals cannot be observed directly, given sensory and computational limits; a proxy is an approximation of the goal that can feasibly be observed, rewarded, and pursued. Consider the abstract goal of reducing a rat population. A continuous rat census is simply not feasible in practice. Even if it were, it would not be possible to incentivize individual agents based on this information, for reasons such as the free-rider problem (Hardin & Cullity, 2020). Therefore, a regulator must operationalize goals based on observable signals that can be incentivized. The Hanoi government chose to measure the number of delivered rat corpses. This *proxy* is *legible* by both the agents and the regulator, it can be used as an incentive, and it is a plausible correlate of the underlying goal. The processing of such a signal constitutes a *regulatory model* (Conant & Ashby, 1970), which captures how proxies are related to agent actions (or traits), and how well these promote the regulatory goal.

Second, regulators are limited in their *ability to predict the future*. For a proxy to perfectly capture a complex goal, the regulatory model would often have to represent all relevant factors and their possible consequences. This may be possible in very simple systems, but in complex biological or social systems a perfect proxy implies a "Laplacian demon": A model that can perfectly predict any future state of the world given any behaviour of an agent (Conant & Ashby, 1970). Absent the ability to perfectly predict the future, a proxy remains prone to unanticipated failure modes in the future.

Third, regulators are limited by the *need to choose*. Decision systems across scales can be associated with *scalar metrics* such as utility, fitness, profit, or impact factor. The reason is that a system tasked with identifying the "better" of any two options in a consistent manner requires options to be rank-ordered. This implies that the options can be described using a scalar metric (see revealed preference approach in economics; Houthakker, 1950; Samuelson, 1938; von Neumann & Morgenstern, 1944). "Consistent" here means that decisions are coherent and not self-contradictory, for instance that they are transitive (if $A > B$, and $B > C$ then $A > C$). Clearly, not all goals can be described in this way (e.g., in socioeconomic systems; Arrow, 1950). But to the degree that consistent regulatory decisions are desirable, a single scalar proxy is implied. It is important to reiterate that this proxy may summarize highly complex, multidimensional information (e.g., multiple contingent input metrics). It may be explicitly constructed (e.g., an objective function in AI, or an academic impact factor), or be implicit, requiring that researchers infer it from an observed set of choices (e.g., utility or fitness) or regulatory system (e.g., voting).

These three limitations shape all regulators, agents, and proxies, because all decision-making systems are practically realized by something physical – and, therefore, something finite. Machine-learning algorithms are limited by computational capacity and training data; peahens and human regulators are limited by the constraints of their eyes and brains. Any proxy reflects the idiosyncrasies, biases, and limitations inherited from the physical system that mediates it.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

3.3. Factors that drive or constrain proxy failure

We propose that a pressure towards proxy failure arises whenever two conditions are met:

- There is regulatory feedback based on the proxy, which has consequences for agents. Agents must be rewarded or selected based on the proxy, which induces optimization of the proxy, either through agent behaviour or through passive evolutionary dynamics. Many commentators emphasize the role of the agent's stakes with respect to proxy performance: The higher the stakes, the higher the pressure for proxy divergence (Muller, 2018; O'Neil, 2016; Smaldino & McElreath, 2016).
- The system is sufficiently complex. What we mean here is that there must be many causal pathways leading to the proxy that are partially independent of the goal (or vice versa). A system may contain proxy-independent actions that lead to the goal, goal-independent actions that produce the proxy, and/or actions that affect both proxy and goal but unequally (Fig. 1).

The first condition leads to proxy optimization. Note that, although in many systems agent behaviour is the outcome of multiple agent-objectives, the present "generalized" account is best served by focusing on only the proxy-objective and considering competing agent-objectives as constraints on proxy-optimization. Regulatory feedback, which may operate through incentives or selection, will result in amplifying the proxy. The second condition implies that there are actions that optimize the proxy but do not necessarily optimize the goal. We posit that, the greater the causal complexity of the system, the lower the probability that the two optimizations perfectly coincide. Although this hypothesis remains to be rigorously tested, a number of system-theoretic considerations support it. Firstly, more complex systems tend to have more "failure modes" (Manheim, 2018). The more complex a system is, the more difficult it will be for a regulator to map *every possible agent action* to the entirety of its consequences. Yet, once the proxy becomes a target, agents will begin a search for actions that efficiently maximize the proxy. In practice, actions that optimize only the proxy often seem to be cheaper for agents than actions that simultaneously optimize the goal (Smaldino & McElreath, 2016; Vann, 2003). Causal complexity may increase the probability that such actions exist. More complex systems may also tend to produce more unstable and nonlinear relations (Bardoscia, Battiston, Caccioli, & Caldarelli, 2017), leading to heavier tails in outcome distributions. This has also been linked to an increased probability of proxy failure (Beale et al., 2020). Intuitively, we can think of an increased probability that some actions exist that *very* efficiently optimize the proxy, but have detrimental (and potentially catastrophic; Bostrom, 2014) effects on the goal. Even if agents are passive and exploration occurs through random sampling or mutation, the system will sooner or later "find" those actions or properties that optimize the proxy in the most efficient manner, irrespective of the consequences for the goal.

Another way to think about this is to note that *opportunities for proxy failure are latent in the countless number of possible future environments*. The regulatory model, and the proxy it gives rise to, can only reflect a finite set of environments. For organisms, this is the environment of evolutionary adaptation (Burnham, 2016); for AI it is the training environment (Hennessy & Goodhart, 2021). By contrast, there are infinite possible future environments (Kauffman, 2019), in which the correlation between proxy and goal may not hold. Because of this imbalance, even random change in a complex environment should be expected to promote proxy divergence. But we can go further: Proxy-based optimization actually biases exploration *towards* such environments. It was precisely the incentive for delivering rat tails that led to rats being bred. An insightful subcategorization of mechanisms underlying this bias has been outlined by Manheim and Garrabrant (2018; see Appendix 2).

Importantly, the presence of a *pressure* towards proxy failure does not guarantee failure: The extent to which a proxy diverges depends on a variety of constraints (which will be elaborated in more detail throughout sects. 4 and 5). For instance, if proxy production diverges sufficiently from goal attainment, the entire system may collapse, as it did in the case of the Hanoi rat massacre (Vann, 2003). In other cases, proxy pressure may lead to inflation, because agent "hacks" to produce more proxy are immediately met by an increase in the proxy level required by the regulator. This appears to have driven school grade inflation (McCoy & Haig, 2020), and it is in fact the natural consequence in any system in which proxy performance matters on a *relative* rather than an *absolute* scale (Frank, 2011). Some proxies may remain informative, even as they are undergoing inflation, such that the pressure to diverge never leads to actual proxy failure (Frank, 2011; Harris, Daon, & Nanjundiah, 2020). In other cases, divergence may be precluded because proxies are directly

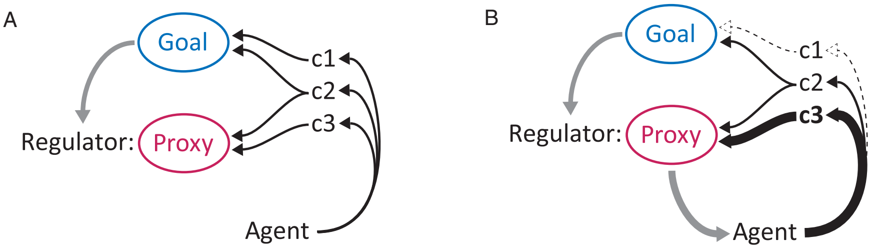

Figure 1. Regulator, goal, agent, proxy, and their potential causal links. Proxy failure can occur when a regulator with a goal uses a proxy to incentivize/select agents. (A) In complex causal networks the causes (arrows) of proxy and goal generally will not perfectly overlap. There may be proxy-independent causes of the goal (c1), goal-independent causes of the proxy (c3), as well as causes of both proxy and goal (c2; note that this subsumes cases in which an additional direct causal link between proxy and goal exists). (B) The regulator makes the proxy a "target" for agents in order to foster (c2). Yet this will tend to induce a shift of actions/properties towards (c3), potentially at the cost of (c1). The causal effects of goals on regulators or proxies on agents are depicted as grey arrows, given that they reflect indirect teleonomic mechanisms such as incentivization or selection. Note that these diagrams are illustrative rather than comprehensive. For instance, the causal diagram of the Hanoi rat massacre would require an "inhibitory" arrow from proxy to goal, as breeding rats directly harmed the goal rather than just diverting resources from it.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

causally linked to the goal, as is assumed for honest (unfakeable) "index" signals in biology. Finally, some decrease (or noise) in the correlation between proxy and goal may be unproblematic for a regulator, depending on its tolerance for such noise (Dawkins & Guilford, 1991).

We will argue that whenever there is some additional selection or optimization pressure that operates on the regulator itself, proxy divergence is constrained. Conversely, in situations where regulators face few constraints, such as in some human social institutions, proxy failure may be persistent and far-reaching (Muller, 2018).

4. Three examples: Neuroscience, economics, and ecology

In this section, we will explore the concepts introduced above in three concrete contexts, neuroscience, economics, and ecology. An overview is shown in Table 3.

4.1. Proxy failure in neuroscience: A new perspective on habit formation and addiction

To pursue objectives, a person implicitly delegates tasks to different parts of their own body and mind. We propose that the dynamics of addiction and other maladaptive habits can be productively understood in terms of proxy failure. Although this framing is novel, its empirical basis is well established (Volkow, Wise, & Baler, 2017). The example reiterates that subpersonal agents and regulators can be susceptible to proxy failure even without intentional hacking or gaming by agents. In doing so, it sheds novel light on a key question in "normative behavioural economics" (Dold & Schubert, 2018), namely why *behaviour* may systematically diverge from *preferences* (Box 2).

The brain can be seen as serving the goal of promoting the organism's fitness, utility, or wellbeing (the present account works for each definition of the goal; see Box 2). The variability of the environment and the diversity of actions available to humans make such goals immensely complex to achieve. To manage this complexity, individual neural systems take up subtasks, such as evaluating the relevance of perceived signals or the consequences of possible actions. For this evaluation, the brain must allocate limited resources such as attention or time, and it must associate perceptual inputs and potential actions with some measure of value. This function is performed by a sophisticated "valuation system," which involves several parts of the limbic system, including medial and orbital prefrontal cortices, the nucleus accumbens, the amygdala, and the hippocampus (Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Levy & Glimcher, 2012). These brain regions are strongly influenced by a neuromodulatory system that plays a key role in neural valuation: the dopamine system (Lewis & Sesack, 1997; Puig, Rose, Schmidt, & Freund, 2014).

Box 2. On revealed vs. normative preferences

Throughout this section, we assume that an addict's *behaviours* can conflict with their *goals*. Behavioural economists refer to this as a divergence of "revealed" and "normative" preferences (Beshears, Choi, Laibson, & Madrian, 2008). It is related to the biological concept of "evolutionary mismatch" (Burnham, 2016), which casts the "normative preference" (goal) as biological fitness. In "evolutionary mismatch," an evolutionary goal – e.g., easily putting on weight to store energy for an uncertain future – lingers as revealed preference that was once also normative. In other words, a change in environment or context led a once adaptive behaviour to become maladaptive. But the behavioural economics concept is broader, in that human goals are not defined as biological fitness but can be any arbitrary notion of, e.g., utility or wellbeing, which may change dynamically (Dold & Schubert, 2018). Importantly, the present account works for either definition of "normative preference," requiring only that revealed and normative preferences are not automatically equated.

It should be noted that the distinction between revealed and normative preferences is not uncontroversial, particularly within economics (Sugden, 2017; Thaler, 2018). For instance, Becker and Murphy (1988) have famously argued that, because an addict takes drugs, this behaviour by definition reflects their preferences. This approach eliminates the possibility of neural proxy failure by assumption (Hodgson, 2003). The issue remains controversial today, as reflected in debates for (Thaler & Sunstein, 2008), or against (Cowen & Dold, 2021; Rizzo & Whitman, 2019), so-called "nudging" as a remedy for what we here cast as neural proxy failure.

Many addictive drugs affect the dopamine system (reviewed in Wise & Robble, 2020). Dopamine is often described as a pleasure chemical, but it is more accurately understood as one of the brain's proxies for "value," broadly construed. The neuromodulator mediates behavioural "wanting" rather than hedonic "liking" (Berridge & Robinson, 2016). Berke (2018) argues that dopamine enables dynamic estimates of whether a limited internal resource such as energy, attention, or time should be expended on a particular stimulus or action. Coddington and Dudman (2019) suggest that dopamine signalling reflects a "scalar estimate of consensus" – that is, a proxy – aggregated from complex inputs from across the brain, to estimate "value." Analogously, the dopamine signal has also been described as approximating "formal economic utility" (Schultz, 2022).

When the dopamine system is functioning appropriately, it contributes to the welfare of the organism. Bursts of dopamine accompany unexpected rewards and other salient events (Bromberg-Martin, Matsumoto, & Hikosaka, 2010), which enable synaptic credit assignment among neural representations of stimuli or courses of action (Coddington & Dudman, 2019; Glimcher, 2011). Over time, dopamine-modulated synaptic weights in motivation-related brain regions (Reynolds, Hyland, & Wickens,

Table 3. Proxy failure in neuroscience, economics, and ecology

Example	Proxy	Goal	Proxy failure	Constraints on proxy failure
Neuroscience	Dopamine signals, synaptic weights	Organism's fitness or wellbeing	• Addictive dynamics • Maladaptive habit formation	• Conscious control (self-control) • Legal and social constraints
Economics	Performance indicators or metrics	Profitability, firm value or "mission"	• Employees maximize indicators at the cost of corporate goals	• Monitoring by managers • Corporate mission • Market selection
Ecology	Cues or signals	Mate or offspring fitness	• Deceptive or runaway signalling	• Compensatory behaviour • Natural selection

Summary of phenomena is described in Section 4.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

2001; Shen, Flajolet, Greengard, & Surmeier, 2008), including the nucleus accumbens, come to serve as proxies for the average value of the corresponding representations. Representations compete for internal resources such as energy, attention, or time (Berke, 2018; Peters et al., 2004). This class of neural competition can be described as "behaviour prioritization," rather than the more specific "decision making," as it spans a spectrum that includes attentional control (Anderson et al., 2016; Beck & Kastner, 2009), exploration (DeYoung, 2013), deliberative weighing of options (Deserno et al., 2015), and unconscious habit formation (Yin & Knowlton, 2006).

Many drugs of addiction interfere with normal dopamine signalling, effectively hacking the neural proxy for value. This "hijacking" of the dopamine reward signal (Schultz, 2022) causes the brain to overvalue drug-related stimuli and actions (Franken, Booij, & van den Brink, 2005; Wise, 2004). Simultaneously, the dopaminergic action of drugs can lead to a decrease in overall dopamine sensitivity, because of well-established biochemical habituation effects (Diana, 2011; Volkow et al., 2017). This leads to a decreased sensitivity to the drug, eliciting the need to seek progressively increasing dosages. At the same time any nondrug-related behaviours become less "valued." The resulting neglect of basic social or bodily needs can lead to severe impairments in health and wellbeing. A similar dopamine-mediated phenomenon can occur in the case of any maladaptive habit, such as excessive unhealthy eating (Small & DiFeliceantonio, 2019; Volkow et al., 2017): The adaptive function of the dopamine system is undermined by the targeted pursuit of dopamine-triggering activities.

Note that the "regulator," here, is the behaviour prioritization system of the addict's brain, and the "agents" are neural representations, which compete for a limited resource: Neuronal credit assignment (Cruz & Paton, 2021). The "regulator" consists of a neural circuit linking the basal ganglia, thalamus, and prefrontal cortex, that mediates competition between synaptic weights and representations (Bullock, Tan, & John, 2009; Mink, 2018; Seo, Lee, & Averbeck, 2012). The brain's "agents" are subpersonal entities, which cannot "game" the system. Instead, the dopamine system modulates synaptic weights, which ultimately represent the "value" of different "representations" of actions or stimuli. Drug use is an action that directly amplifies these proxies independent of their usual relationship to wider goals. Therefore, use-related synaptic patterns are strengthened, leading to increased drug-seeking.

The key limitations necessitating a proxy system are legibility, prediction, and the need to choose (sect. 3.2). Dopaminergic drugs mediate changes to synapses that are "legible" to the brain's valuation system: They manipulate the "currency" that mediates ordinary decisions. Natural selection cannot possibly anticipate all possible experiences and chemicals that can trigger elevated dopamine, reflecting the limits of prediction in a complex system. Similarly, the pattern-recognition systems in the brain cannot anticipate all possible consequences of the organism's actions. The need to choose implies that the brain must rank-order possible actions at any given moment.

But neural proxy failure is nevertheless contingent on numerous additional factors. A propensity for addiction has been linked to factors in childhood and adolescence (Leyton & Vezina, 2014). More generally, most individuals consciously constrain behavioural impulses to use drugs or engage in other potentially maladaptive behaviours on a day-to-day basis. Indeed, the conscious shaping of habits and preferences appears crucial to a self-determined life (Dold & Schubert, 2018). Such individual efforts are embedded into legal and social constraints that can enable or discourage addiction and related behaviours (Braganza, 2022a).

In summary, addiction and maladaptive habit formation can be viewed as neural instances of proxy failure. To our knowledge, this is the first time that a neural phenomenon has been described in terms of Goodhart-like dynamics. The proxy failure perspective suggests that the risks of addiction and maladaptive habit formation may be natural consequences of pursuing goals via neural representations of (i.e., proxies for) value. If this is the case, a "brain disease" framing of addiction may contribute to excessive medicalization (Satel & Lilienfeld, 2014) at the expense of exploring structural changes at broader environmental, social, or economic levels (Heather, 2017). Later, we will explore a case in which humans appear to harness proxy divergence within themselves in order to support their conscious goals (sect. 5.3).

4.2. Proxy failure in economics: Performance metrics and indicatorism

The disciplines in which proxy failure may have been studied most intensely, albeit under different names, are management and economics (Holmström, 2017; Kerr, 1975; van der Kolk, 2022). Consider, for instance, the use of key performance indicators (KPIs). Price and Clark (2009) state in unequivocal terms that: "people who are measured on KPIs have an incentive to play games. […] The indicator ceases to indicate." A prominent early example was Lincoln Electric's plan to incentivize typists in its secretarial pool through a piece-rate for each key stroked (The Lincoln Electric Company, 1975). This apparently led to typists continuously tapping the same key during lunch hours (Baker, 2002). Many additional entertaining examples are collected in Stephen Kerr's management classic: "On the folly of rewarding A, while hoping for B" (Kerr, 1975). More recent examples include the Wells Fargo scandal, where excessive incentives led to fraudulent "cross-selling" (opening additional accounts for customers without their knowledge and against their interest; Tayan, 2019). Many more empirical examples have been described in management and accounting research (Bonner & Sprinkle, 2002; Bryar & Carr, 2021; Franco-Santos & Otley, 2018) – for example, concerning CEO pay (Bénabou & Tirole, 2016; Edmans, Fang, & Lewellen, 2017).

Inspired by such empirical observations, economic theorists have developed a range of mathematical equilibrium models (Box 3), which describe proxy failure within what is known as the principal–agent framework (Aghion & Tirole, 1997; Baker,

Box 3. Proxy failure and equilibrium models

Formal economic as well as ecological models often rely on equilibrium analysis, implicitly assuming an unchanging world. This is partially owed to prevalent methodological traditions that are focused on analytic models. Proxy failure in such a framework is by definition *static*, where divergence, if present, has progressed to its final level. This equilibrium assumption is intricately linked to ongoing controversies on proxy failure across disciplines (see, e.g., Box 5).

Disequilibrium research recognizes that many systems exist outside equilibrium (Wilson & Kirman, 2016). Examples of proxy failure may be observable within corporations because markets have not *yet* had time to offer corrective feedback to *novel* forms of "hacking." More generally, a dynamic perspective raises the question of whether equilibrium analysis is sufficient to capture proxy failure, which may be best understood as an "out-of-equilibrium" phenomenon.

2002; Bénabou & Tirole, 2016; Holmström, 2017). The task of the principal or "regulator" is cast as the design of an "optimal incentive contract," given noise or the potential for active "distortion" (Baker, 2002; Hennessy & Goodhart, 2021). In other words, the principal must translate observable agent outputs into regulatory feedback (e.g., monetary incentives or promotions). The recurring result is that the potential for distortion implies a *weaker* optimal incentive strength (Baker, 2002). In other words, whenever proxies do not perfectly capture the goal, a reduction in the reward for the proxy is warranted. In fact, a particularly important question in the early literature was why pure fixed wage contracts (i.e., contracts without performance-contingent pay) are so common in real businesses (Holmström, 2017). Why do companies not generally, or at least more frequently, base wages on measured performance? The answer is: The risk of proxy failure.

Another focus of the economic literature is the role of proxies in large hierarchical organizations. For instance, we can think of KPIs as the communication interface between general management and individual departments. Aghion and Tirole (1997) develop a theory of delegation of authority in hierarchical corporate contexts, exploring the complex roles of information and "incentive congruence" in determining how proxies can be gainfully used. More generally, proxies appear to reflect the division of labour in complex organizations or multistep processes, which may occur on different timescales. Individual employees, particularly in sales and marketing roles, are typically incentivized for achieving short-term objectives such as signing a contract with a customer, or making contact with a potential buyer. Whole departments can then be measured on the aggregate numbers achieved by their employees. Each department or subdivision must regulate its own subdivisions, necessitating further proxy measures. The use of proxies at each level of such a hierarchy naturally entails the risk of proxy failure, which indeed has been extensively documented. For instance, Tardieu, Daly, Esteban-Lauzán, Hall, and Miller (2020) explore proxy failure in the context of "parochial" KPIs: Those measured by individual departments that are imperfectly related to wider business outcomes. Ittner, Larcker, and Meyer (1997) provide an example from a large bank: "A system put in place to provide incentives for branch managers to increase customer satisfaction quickly began to reward some of the most unprofitable branches" (Baker, 2002).

In all these examples proxies diverge from the underlying goals of an organization. But proxies are nevertheless useful, indeed unavoidable, in practice. The informational factors necessitating proxy use by organizations are special cases of the three limits that necessitate proxies (sect. 3.2). Agents in a complex organization often have insufficient knowledge of the top-level organizational goal and how to promote it – the "legibility" limit. Proxies are expedient means of communicating the goal through department hierarchies to coordinate agent behaviour. The prediction limit describes that neither employers nor employees can foresee all the consequences of a given behaviour for the organization. The choice limit is reflected in the frequent evaluation of infinite possible employee actions by single scalar metrics such as KPIs. In addition, psychology is at play: Without direct incentives, which are (i) close in time, (ii) relatively certain, and (iii) directly personally relevant, employees may have little motivation to contribute to an organizational goal (Jones & Rachlin, 2009; Kobayashi & Schultz, 2008; Strombach et al., 2015).

Another core insight of the economics and business literature concerns a fundamental constraint on proxy failure. If we posit a corporation's goal as competitive success in the market, proxy failure within businesses is naturally constrained by market selection (Braganza, 2022b). If proxies designed to regulate employees diverge too far from the corporation's goal, profit and share prices may plummet – causing behaviour change. In the worst-case scenario, the whole organization may go bankrupt. Market selection thus provides a hard constraint on proxy failure within a business. Whether implemented by savvy managers or market forces, the constraint controls regulatory models themselves. Because of this, the economic literature has tended to view the very existence of a given practice in real corporations (e.g., fixed wages) as evidence of its economic optimality (e.g., Baker, 2002; Holmström, 1979, 2017).

This insight also explains why proxy failure is likely even more relevant in nonmarket-based economies, where the hard constraint of market selection is missing. An apocryphal anecdote (Shaffer, 1963) tells of a Soviet factory rewarded by the Central Committee for the *number of nails* it produced, resulting in the output of millions of flimsy, useless nails. When the proxy was changed to instead reward the *weight* produced, the factory retooled to manufacture one single immense nail. Without the market constraint, proxy failure can remain unmitigated. Specifically, market proxies (e.g., profitability) ensure that consumers must actually find a product useful by measuring whether they buy it. However, the assumption that market proxies thus *automatically* serve our societal goals is also clearly problematic (Box 4).

Finally, many scholars have pointed out that humans are to some degree purpose driven, that is, not solely interested in extrinsic (proxy) incentives and that extrinsic incentives can indeed undermine worker morale and human wellbeing more generally (Muller, 2018; van der Kolk, 2022). This suggests an approach to mitigating proxy failure that is unique to human contexts: The promotion of intrinsic incentives towards the goal (Mayer, 2021; Pearce, 1982). Some firms advertise "social objectives," such as to "Accelerate the world's transition to sustainable

Box 4. On market selection and social welfare

Important proxies in economics are *price* and *profit* (Friedman & Oren, 1995; Mas-Colell, Whinston, & Green, 1995). In the standard economic view, markets are collections of players who regulate each other through price competition: Buyers (as regulators) seek out the lowest marginal price/utility ratio, using price as a proxy to decide the best supplier (agent); conversely, sellers take on the role of regulator and supply to whichever agent will pay the highest price. A secondary implication is that *profit* functions as a proxy for welfare increases, systematically directing economic resources to where they will most efficiently enhance welfare. In theory, this results in an equilibrium where all achieve their goals and total consumption directly tracks total welfare (Fellner & Goehmann, 2020).

In practice, many proxy failures are seen. Economists think of such cases as "market failures." A prominent example is when the interaction between buyers and sellers adversely affects third parties (i.e., entails a so-called "negative externality"), for instance because of environmental degradation or the risk of catastrophic climate change (Fremstad & Paul, 2022; Wiedmann, Lenzen, Keyßer, & Steinberger, 2020). These phenomena imply that some welfare costs are not accounted for in the price and profit proxies leading to "decoupling" from the underlying goal (Kelly & Snower, 2021; Mayer, 2021; Raworth, 2018). One canonical way to address such proxy failure is by leveraging "Pigouvian taxes" (e.g., a tax on CO_2 emissions), which can "internalize externalities" into market prices, thus realigning the market-level proxies with the underlying societal goal (Masur & Posner, 2015).

Another mechanism that potentially decouples economic proxies from their underlying goal is "persuasive advertising" (Box 8).

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

energy” for Tesla, to provide both guidance and motivation to employees, beyond their material incentives. This could mitigate proxy failure by directly changing agents’ utility, as modelled in the multitasking literature (Braganza, 2022b; Holmström, 2017). It may also provide a second-order protective function by promoting employee vigilance towards the possibility of unanticipated avenues for proxy failure. Mayer (2021) suggests that an explicit corporate purpose, combining both profitability and social/ecological responsibility, can mitigate proxy failure both within a firm and in the larger societal context.

In summary, economics and business research reveals key insights into proxies and their potential for failure. From theories of contract structure and measurement, to the theory of prices as a signal for optimal resource allocation (Friedman & Oren, 1995), to political debate about the effectiveness of the profit motive – choosing and evaluating proxies is a core activity of economists. The present perspective places this literature within a larger framework, suggesting that the mechanisms explored therein may inform, and be informed by, mechanisms studied in other disciplines.

4.3. Proxy failure in ecology: Sexual selection, nested goals, and countervailing forces

Recently, McCoy and Haig (2020) have argued that proxy failure occurs in ecological systems in a manner analogous to many social systems. Here, we introduce and expand their argument and highlight the intriguing similarities to the case of economics.

Proxy failure provides a useful framework to understand evolution. For instance, peahens regulate peacocks via mate choice and mothers regulate offspring via embryo selection. The proxy in such cases is any trait or “signal” expressed by the regulated agent that affects regulation, such as a peacock’s tail or an embryo’s ability to produce hormones. Individual organisms can respond to such pressures as humans do, by deliberately increasing production of the proxy; for example, males of many species prefer to display near less attractive rivals, apparently so that they can themselves appear better (Gasparini, Serena, & Pilastro, 2013). But deliberation is not necessary – animals with greater proxy values may also simply be selected and pass on proxy-producing genes. A discrepancy between a signal and the information it was originally selected to convey, that is, proxy failure, is often termed *deception* (without implying intentionality). It seems undisputable that both honest and deceptive signalling are widespread in ecological systems (Dawkins & Guilford, 1991; Krebs & Dawkins, 1984; McCoy & Haig, 2020; Wickler, 1965). However, the question of whether stable ecological signals can safely be considered “honest” and “adaptive” has been a subject of controversy since Darwin (Box 5).

Box 5. On “honest” signalling or “adaptationism”

In “The Evolution of Beauty” Richard Prum (2017) reviews the contested history of “honest signalling” theories (also see Prum, 2012). Adaptationists, whom Prum and others (e.g., Haig, 2020; Lynch, 2007) identify as the mainstream of evolutionary biologists, assume that biological signals should be assumed “honest” or “reliable” indicators of underlying fitness unless proven otherwise (Zahavi et al., 1993). In this view, natural selection is *the* key force in evolution, and any trait or signal that exists can safely be assumed “adaptive” or “reliable” at equilibrium. The implication is that *proxy failure* and unreliable signalling more generally are of negligible importance, because (at equilibrium) they would have been eliminated by natural selection. One prominent account termed “costly signalling” theory (Spence, 1973; Zahavi, 1975) suggests that seemingly wasteful signals reliably communicate fitness because fitter individuals can afford more of them. Although many in ecology (and economics; Connelly et al., 2011) accept honest signalling theories such as costly signalling as a “scientific principle,” others view it as an “erroneous hypothesis” (Penn & Számadó, 2020; Számadó & Penn, 2018). Interestingly, conclusions often seem to hinge on which parts of a system are tacitly assumed to be at equilibrium (Penn & Számadó, 2020).

By contrast, Darwin’s theory of sexual selection suggested that “aesthetic” traits could be selected and become stable *independent* of their implications for “fitness,” i.e., their “adaptive value.” Later scholars have coined the term “runaway selection” for such processes (Fisher, 1930; Prum, 2010; Rendell et al., 2011), suggesting they may underlie much of the diversity of display traits in the animal kingdom (Prum, 2017; also see sect. 5.1 and Box 6). From the “adaptationist” perspective, runaway signals must be viewed as proxy failure because signals do not reliably indicate fitness.

Sexual selection in colourful birds and fish is a useful case study for proxy failure (McCoy et al., 2021). Many female birds and fish choose a mate based on a proxy signal of quality. A famous proxy is carotenoid-based pigmentation: These red–orange–yellow colours are often considered “honest signals” of fitness, either because they incur a cost (Spence, 1973; Zahavi, 1975) or are causally linked to metabolic processes (Maynard-Smith & Harper, 2003; Weaver, Santos, Tucker, Wilson, & Hill, 2018). It is thought that female birds and fish prefer *redder* males because this trait reliably (i.e., honestly) communicates their fitness. However, some male fish and birds have evolved strategies to maximize the proxy without increasing their carotenoid pigments. Some male fish have evolved the ability to synthesize alternative (potentially cheaper) pigments (pteridines; Grether, Hudon, & Endler, 2001); some male birds have evolved light-harnessing structures that make their feathers appear more colourful without requiring more pigment (McCoy et al., 2021). Females have sensory limits; they cannot always tell *what* produces the red colour. In other words, the evidence suggests that carotenoid-based signalling is not always honest – male birds and fish can “game” the system, driving proxy failure.

Other apparent instances of proxy failure appear in mother–embryo conflict. Babies are costly; therefore, mothers assess embryo quality early and automatically abort less-fit embryos based on proxy signals. Embryos are incentivized to “hack” this system and ensure their own survival. For instance, marsupial infants are born mostly helpless and must find and latch on to a maternal teat to survive. However, some marsupials birth more offspring than they have nipples available, so the infants must literally race to their mother’s nipples. Slower individuals are doomed (Flynn, 1922; Hill & Hill, 1955). For example, Hartman (1920) observed 18 neonates of *Didelphis virginiana* reaching the pouch and competing for only 13 teats. Mothers seem to screen infants for quality through this race to the teat; speed is a proxy of quality. In response, infants are born with exceptionally strong forearms – they overinvest in performance at the maternal test. A weaker infant who invested their developmental energy into forearms will win against a healthier rival who invested their developmental energy equally across body parts.

Plants similarly show instances of proxy failure in embryo selection. In many plants, far more initial embryos are produced than ultimately survive, because of various stages of maternal screening and selective abortion, but embryos have evolved many strategies to boost their chances of being selected (Shaanker, Ganeshaiah, & Bawa, 1988)**.** For example, the plant *Caesalpinia pulcherrima* produces seed pods that have between 1 and 10 seeds per pod, and the mother selectively aborts pods with fewer seeds. However, some of these “less fit” pods evade abortion attempts – and then absorb their siblings, acting in a manner contrary to the evolutionary interests of the mother. Similarly, embryos in the woody tree *Syzygium cuminii* secrete factors to promote abortion of their peers – potentially a response to maternal selection for healthy embryos (Arathi, Ganeshaiah, Shaanker, & Hegde, 1996). For a thorough treatment, see Willson and Burley (1983). Another example of proxy failure in embryo selection in humans, other primates, and horses is presented in Section 5.1.

So what constrains proxy failure in ecology – which forces can keep signals honest (Andersson & Simmons, 2006; Connelly, Certo, Ireland, & Reutzel, 2011; Harris et al., 2020; Weaver et al., 2018)? We have already noted the controversial nature of questions regarding when or even if ecological signals are honest (Box 5). Here we sidestep such controversies by simply reviewing proposed mechanisms without attempting to claim a general validity (or falsity) of any specific mechanism. In general, the natural selection of regulators should promote "honest signals" (Maynard-Smith & Harper, 2003). Such selection should (i) constrain divergence where it occurs and (ii) systematically lead to the identification of proxies that are more resistant to failure. First, even though we have outlined cases in which presumed "honest signals" were gamed, this does not imply that some proxies may not be more resistant to gaming than others, for example, because they are more tightly causally linked to the actual goal (so-called "index signals"; Maynard-Smith & Harper, 2003; Weaver et al., 2018). Another proposed mechanism to keep signals honest, which has been highly influential across ecology and economics (Connelly et al., 2011), is "costly signalling" or the "handicap principle" (Zahavi, 1975; Box 5). It is the idea that, if the cost of a signal is negatively correlated with fitness, then fitter individuals can afford to produce more signal. Regardless of mechanism, selection should favour those individuals that do not unnecessarily fall prey to deceptive signals. This in turn should lead to the selection of more honest signals whenever possible.

Another intriguing parallel to the business literature is that life is often organized in nested hierarchies (Ellis & Kopel, 2019; Haig, 2020; Kirchhoff et al., 2018; Okasha, 2006). Goals at one level of natural selection can become proxies at another level (Farnsworth, Albantakis, & Caruso, 2017). Species compete within ecosystems; individuals compete within species; cells compete within individuals; genes compete within genomes; and so on (Okasha, 2006). At the highest level, natural selection regulates all living creatures (agents) according to how well they pass their genetic material on to future generations (the goal – see Appendix 1 for discussion of this choice of word). This genetic goal is operationalized through many simpler behavioural proxies: Most creatures are "wired" to survive predation, find food, attract a mate, invest in healthy offspring, and more. These proxy tasks are not themselves important except insofar as they contribute to the goal of passing on genetic material. Each of these proxies has become a goal, itself operationalized by further proxies.

This nested hierarchy offers many future areas of research into proxy failure in evolution. Consider, for example, "selfish" genetic elements (Ågren & Clark, 2018; Haig, 2020) that disproportionately insert themselves into the genome of an organism's offspring; they have "hacked" the organismal proxy of reproduction by jumping up a level in the hierarchy to the actual goal (transmitting genetic information). Certain weeds have evolved to mimic crop plants through a proxy divergence process known as Vavilovian mimicry (Vavilov, 1922), requiring a "model" (the crop), "mimic" (weed), and "operator" (human or herbicide) (McElroy, 2014). Similarly, signalling between predators and prey, plants and animals, cleaner fish and their clients, and more, is likely all subject to proxy failure and its consequences.

In summary, proxies and proxy failure appear to be as common and well researched in ecological systems as in economic systems, albeit characterized using different terms, such as deceptive or "runaway" signalling. Although the regulatory feedback mechanism is often selection rather than incentivization, the similarities in the resulting dynamics are striking (McCoy & Haig, 2020). Yet it should also be emphasized that this does not negate the important differences between biological and human contexts: It seems neither accurate nor desirable to reduce individual and social human goals to *biological fitness*.

5. Consequences of proxy failure: The proxy treadmill, the proxy cascade, and proxy appropriation

In the previous sections we have defined proxy failure and given examples from diverse fields. Here, we will describe three important consequences of the phenomenon. In each case proxies, goals, regulators, or agents, across multiple systems interact. An overview is shown in Table 4.

5.1. *The proxy treadmill: When agent and regulator race to arms*

McCoy and Haig (2020) describe a second-order dynamic arising from proxy failure, termed the "proxy treadmill" (Fig. 2). The proxy treadmill describes an arms race between agent and regulator, leading to ratcheting cycles of hacking and counter-hacking. This leads to the emergence of three hallmark phenomena, namely (i) instability, (ii) escalation, and (iii) elaboration. We argue below that these phenomena promote the seemingly inexorable growth of complexity in both biological and social systems.

- *Instability* arises in a system when an arms race between regulator and agents prevents stable equilibria from being attained. Optimization implies that agents will never stop "searching" for new ways to hack a signal, and given the complexity of ecological systems, they are likely to find one eventually, triggering a new cycle of counter-hacking by the regulator.

Table 4. Overview of consequences of proxy failure

Phenomenon	Description	Consequences
Proxy treadmill	An arms race between agent and regulator to hack and counter-hack a proxy emerges.	• Instability of signalling systems • Inflation/escalation of signals • Signal elaboration, increasing complexity
Proxy cascade	In a nested hierarchical system, a higher-level proxy constrains divergence of lower-level proxies.	• The system as a whole can pursue the higher-level goal and dynamically adapt to lower-level proxy failure
Proxy appropriation	In a nested hierarchical system, goals at one level are served by harnessing proxy failure at other levels.	• The goals at one level (e.g., human individuals) can be prioritized and guide overall system behaviour

Summary of phenomena is described in Section 5.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

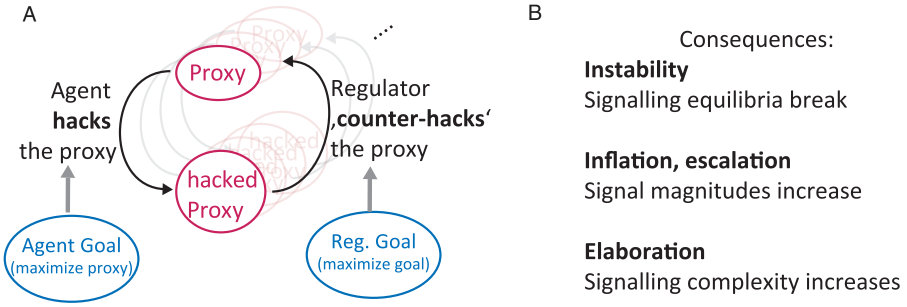

Figure 2. The proxy treadmill. Illustration of the proxy treadmill and its consequences. (A) An agent "hacks" a proxy by, say, finding cheaper ways to make it. The regulator responds by "counter-hacking," for example, requiring higher levels of the proxy, or requiring new proxy attributes. Each pursues their own goal (which for the agent, here, is reduced to wanting to maximize the proxy). (B) This results in instability, escalation, and elaboration of the signalling system.

- *Escalation* describes a quantitative manifestation of instability, whereby agents find novel ways to produce more of the same proxy, and regulators respond by requiring higher proxy levels. It can account for the size of peacock tails, the inflation of embryonic hormone production, the inflation of test scores in education, or publication counts in academia.
- *Elaboration* describes a qualitative manifestation of instability, whereby proxy failure continuously drives the modification of given proxies, or the addition of new proxies. It helps characterize the evolutionary elaboration of the peacock's tail into complex shapes, the chemical modification and addition of hormones secreted by embryos, and the addition of novel criteria in education (extra-curriculars) or academic competition (alt-metrics).

McCoy and Haig (2020) introduce the proxy treadmill and its three hallmarks in the context of mother–embryo conflict, reviewing traces it has left in the genomes of both primates and horses. A signal originally used by the mother as a proxy for embryo viability during early pregnancy, luteinizing hormone, was hijacked by the embryo to improve its own chances of survival: Embryos produce a hormone that mimics luteinizing hormone called chorionic gonadotropin or CG (Casarini, Santi, Brigante, & Simoni, 2018). Several successive rounds of hacking and counter-hacking appear to have ensued, leading to an *escalation* to a present in which "heroic" amounts of hormone are produced by embryos (Zeleznik, 1998) and required by mothers (McCoy & Haig, 2020). Simultaneously, *elaboration* of the signal occurred, as embryos accumulated mutations that changed the structure of the hormone, for instance increasing the half-life and thereby the concentration of the hormone in circulation (Henke & Gromoll, 2008). This illustrates how *instability*, *escalation*, and *elaboration* have shaped the present signalling system. Analogous processes occur when embryos evolve new "mimics" of proxy signals or mothers interpret existing signals in new ways (see Haig, 1993; McCoy & Haig, 2020). The authors liken this to the addition of novel selection criteria in job interviews, after grades have inflated to a degree of relative uninformativeness.

A key tenet of the proxy treadmill is that uninformative proxies may sometimes linger for a variety of reasons, despite the pressures of natural selection on both the regulator and the agents. For instance, regulators may keep observing relatively uninformative proxies because of a *first-mover disadvantage*, which is closely related to Fisherian runaway signalling (Fisher, 1930; McCoy & Haig, 2020; Prum, 2010; Box 6). Fisher argued that any peahen with a preference for peacocks with small tails will have sons who are unattractive to other peahens. In other words, large tail sizes could stabilize, not because they signal fitness, but because the male trait and the female preference for that trait are jointly inherited. By this mechanism, competitive selection can lock in a real selective advantage for an otherwise uninformative, or even a fitness-decreasing, proxy. McCoy and Haig note the similarity with Princeton University's unilateral attempt to halt grade inflation (Stroebe, 2016): No other universities followed suit, students began ranking Princeton lower, and Princeton eventually abandoned the policy (Finefter-Rosenbluh & Levinson, 2015). Other reasons why uninformative proxies may linger are simple slack in the selection of regulators or that it can be economical for regulators to employ even poor proxies, if the cost of acquiring more accurate information is high (see Dawkins & Guilford, 1991).

The proxy treadmill bears striking similarities to what might be called the "academic proxy treadmill" (Biagioli & Lippman, 2020), and the "bureaucratic proxy treadmill" (Akerlof & Shiller, 2015). Biagioli and Lippman (2020) describe how in academia, "new metrics and indicators are being constantly introduced or modified often in response to the perceived need to adjust or improve their accuracy and fairness. They carry the seed of their never-ending tuning and hacking, as each new metric or new tuning of an old one can be subsequently manipulated in different ways." Similar mechanisms of continuous hacking and counter-hacking appear to occur when a government regulates corporations, which find ways to hack the regulations, triggering novel regulations (Akerlof & Shiller, 2015). The predictable results, analogous to biological cases, are continuously expanding and ever more complex bureaucracies (Muller, 2018).

Together, this suggests that the proxy treadmill may act as a driver of complexity in both biological and social systems. Some proxies may rapidly inflate and then completely fail, as was the case for the Hanoi rat massacre. Other proxies may inflate and

Box 6. Relation between the proxy treadmill and Fisherian "runaway"

The proxy treadmill generalizes a prominent model of sexual selection, namely Fisherian runaway (Fisher, 1930; Lüpold et al., 2016; Prum, 2010). Fisherian runaway is one possible instance of a proxy treadmill in the domain of *sexual selection*, which rests on the genetic association between a trait and a preference because of coinheritance. A related mechanism is "runaway cultural niche construction," where the association of trait and selection mechanism (the niche) occurs because of geographic colocalization (Rendell et al., 2011). Both models answer the question of why signals might escalate (i.e., run away), positing specific association mechanisms. Modelling papers have also shown that Fisherian runaway may promote trait elaboration (Iwasa & Pomiankowski, 1995). The proxy treadmill is agnostic about the association mechanism and even whether one exists. Instead, all it requires is a conflict of interest between low-quality males and choosy females: In its minimal form the proxy treadmill reflects simply the continuous breakdown and replacement of signals.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

then remain stable, even though they have become less informative, such as inflated grades (McCoy & Haig, 2020), publication metrics (Biagioli & Lippman, 2020), or "aesthetic signals" (Prum, 2012). Still other proxies may be impossible to inflate because they are causally linked to the goal (index signals) or may remain informative even while inflating (costly signals; Harris et al., 2020; Zahavi, Butlin, Guilford, & Krebs, 1993). Crucially, any change in a proxy or incorporation of a new proxy by the regulator gives rise to novel affordances for the agent, which can lead to new rounds of hacking and counter measures. The result is an ecology of signals, constantly being renegotiated, but tending to increase in complexity and elaboration over time. Note that we do not mean to imply it is the only driver of complexity; as for example in biological systems numerous additional forces have been studied (Box 7).

In summary, proxy failure may lead to a second-order dynamic: The proxy treadmill, an arms race of hacking and counter-hacking that causes instability, inflation, and elaboration. This phenomenon again highlights the question of whether equilibrium analyses may systematically obscure some of the more interesting behaviours of biological and social systems: Signalling systems may remain far from equilibrium because of high levels of instability, and equilibria that are reached may be transient. Proxy failure and the proxy treadmill may be key drivers of complexity in both biological and social systems.

5.2. The proxy cascade: When high-level proxies constrain lower-level proxy failure

Throughout this paper, we have repeatedly noted that proxy-based decision systems are embedded in larger hierarchies, where high-level goals may constrain or shape low-level goals (Heylighen & Joslyn, 2001). For instance, markets contain corporations, which contain departments, which contain individuals (Aghion & Tirole, 1997). Individuals are themselves nested hierarchical systems, containing organs such as the brain which in turn are organized in subsystems, down to the level of cells and organelles (Farnsworth et al., 2017). How, then, might proxies and goals interact across multiple levels? Here, we argue that high-level regulation can constrain and control proxy failure at all lower levels – a cascade of regulation.

Box 7. Drivers of biological complexity

Importantly, the proxy treadmill is just one driver of biological complexity. Many additional forces, both adaptive and nonadaptive (Lynch, 2007), likely play key roles: The accumulation of neutral mutations (Gray, Lukeš, Archibald, Keeling, & Doolittle, 2010), the evolution of mechanisms to suppress conflict at lower levels (Buss, 1988), and other evolutionary arms races. The relation between the proxy treadmill and other evolutionary arms races (e.g., predator–prey or host–parasite) is particularly close, and so merits disambiguation: The proxy treadmill explicitly concerns a signal and its interpretation, which need not be present for other arms races.

An important point to reiterate before we proceed is that the identification of proxy failure is inseparable from the definition of a goal at a specific level. In a hierarchy, the proxy of a "parent" regulator typically becomes the goal for a "child" agent. The failure claim is thus specific to the goal of the regulator but not to that of the agent. We emphasize this point, because considering proxy failure across levels requires particular attention to (i) goals at multiple levels and (ii) the specific goal that underlies the claim of proxy failure.

How can regulation cascade down a sequence of proxies in a nested hierarchy, effectively allowing the pursuit of integrated, high-level goals? Consider a corporation in which management has created a proxy (*sales revenue*) to promote its goal (say, *profitability*; Fig. 3). The sales department adopts sales revenue as their goal, and therefore creates a proxy (the *number of telephone calls made*) to motivate the outputs of its salespeople. The proxy of each higher level becomes the goal of the lower level. Proxy failure can occur at any level. But importantly, the corporation itself is subject to regulation at a still higher level, namely market selection.

First, assume proxy failure occurs at the department level: The sales department artificially inflates *sales revenue* numbers, perhaps by an accounting trick. If such department-level proxy failure undermines firm profitability (the firm-level goal) to a large enough extent, then bankruptcy might ensue. In practice, such an outcome may be mitigated by corporate management, which is thus efficiently incentivized to constrain department-level proxy failure. Regardless of the mechanism, market selection provides a higher-level constraint on the degree of proxy divergence at the department level.

Second, assume proxy failure occurs at the subdepartment level. Recall that the sales department chose the *number of calls made* as its proxy for the goal of increased revenue. Now employees have the goal to maximize the number of calls, but they may do this in ways that do not lead to increased sales. Perhaps employees purposely keep calls short to increase total number, even though they would be more likely to make a sale if they stayed on the line longer with potential clients. This would undermine the department-level goal as well as the corporate-level goal: More calls, lower sales revenue, less profit. Would this lower-level

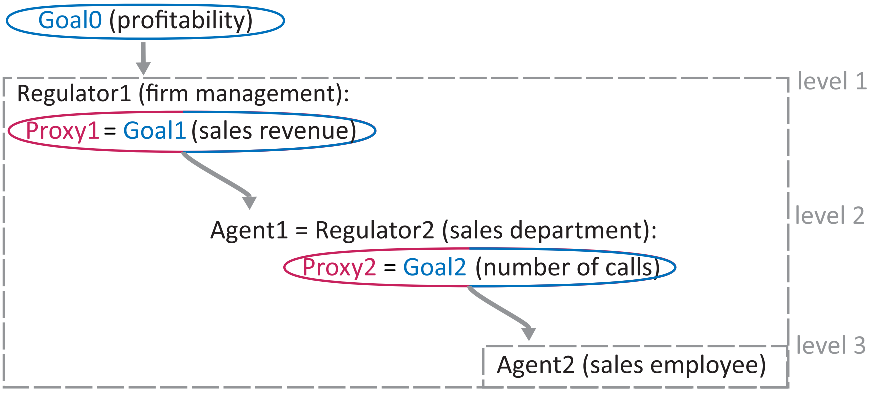
Goal0 (profitability)
level 1
Regulator1 (firm management):
Proxy1 = Goal1 (sales revenue)
level 2
Agent1 = Regulator2 (sales department):
Proxy2 = Goal2 (number of calls)
level 3
Agent2 (sales employee)

Figure 3. The proxy cascade. Proxies often regulate the interaction between nested modules of complex hierarchical systems, such as corporations. Goals of lower levels tend to be proxies used by higher levels to regulate and coordinate those lower levels. The proxy cascade describes how a higher-level constraint cascades down the levels, constraining proxy failure at all lower levels.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

proxy failure also be constrained? It arguably would. First, if the department-level proxy is constrained to function reasonably well, then department managers are efficiently incentivized to mitigate lower-level proxy failure. Second, if, for whatever reason, subdepartment-level proxies do excessively diverge, the firm will go bankrupt just like a corporation with excessively divergent proxies at the department level.

Therefore, wherever proxy failure occurs in a hierarchy, it seems to be constrained by the highest-level proxy – what we term a *proxy cascade*. In the words of Heylighen and Joslyn (2001), "higher-level negative feedbacks typically constrain the growth of lower-level positive feedbacks." Informational limits (sect. 3.2) will still produce the pressure towards proxy failure and runaway dynamics at each level, but regulation of the regulators causes a *counter pressure which cascades down the levels*. At each level, this may lead to dynamic phenomena such as the proxy treadmill, the identification of more stable or honest proxies, or a mix of such phenomena. Overall, the result is a noisy self-organization of the cascade of proxies, ultimately allowing the integrated pursuit of a high-level goal.

The proxy cascade also seems to capture adaptationist (see Box 5) aspects of the interaction between natural selection and sexual selection. For example, female birds choose males based on how beautiful the males are – a proxy of their quality as a mate. But if the male develops ornaments that are too cumbersome or burdensome, the male (and his offspring) cannot survive the rigours of natural selection. This is not merely theoretical; Prum (2017) describes "evolutionary decadence," by which female aesthetic choice shapes outrageous ornaments in certain bird species – potentially increasing extinction risk. Beyond mate choice, another instance of the proxy cascade concerns cooperative proxies for intimately associated organisms (such as hosts and symbionts); when cooperation is not properly enforced, or when proxies for cooperation fail, species face extinction risk (Ågren, Davies, & Foster, 2019); for example, *Wolbachi* endosymbionts threaten survival of populations of butterfly by altering sex ratios (Dyson & Hurst, 2004). In this manner, natural selection may exert a constraining hand over proxy failure at lower levels in ecology. We offer these as initial ideas of proxy cascades in biology, but note that substantial further theoretical and empirical research is needed.

The proxy cascade suggests that high-level regulation can constrain lower-level proxy failure. It explains why high-level selection systems, such as natural selection or the market, can so efficiently coordinate complex structures towards coherent high-level goals. Further, the cascade of proxies suggests another possible opportunity for our pursuit of human goals – and that is the subject of the next section.

5.3. Proxy appropriation: When high-level goals hack low-level proxies

Finally, we introduce the novel idea that higher-level systems may harness or "appropriate" proxy failure at lower levels to achieve their high-level goals. We illustrate this using the example of artificial sweeteners.

Humans sense sweetness via the oral T1 receptor system, which employs a molecular "lock and key" mechanism (DuBois, 2016; Li et al., 2002). The "goal" of such taste receptors is to enable the organism to base food choices on estimates of the nutritional value of ingested foods. The specific mechanism by which "assessment" takes place resides in the intricate molecular structure of the receptor dimer (the lock; Perez-Aguilar, Kang, Zhang, & Zhou, 2019), which was shaped by evolution to "identify" valuable substrates (the keys). If "lock" and "key" sufficiently match, the molecules bind and T1 receptors initiate a complex signalling cascade, which ultimately feeds into a value estimate of the neural reward system (Breslin, 2013). Sweetness, as assessed by T1 receptor binding probability, can thus be understood as a proxy for nutritional value. All organisms, from bacteria to humans, rely on similar molecular assessment mechanisms to "infer" the nutritional value of molecular substrates, in order to guide ingestion decisions towards the goals of survival and homoeostasis (Jeckelmann & Erni, 2020).

If receptor binding probability can be considered a "molecular proxy," then what constitutes hacking? Some molecules might mimic the specific biophysical properties of the target molecules leading to "non-specific" binding (Frutiger et al., 2021). The artificial sweetener aspartame binds to the T1 receptor, eliciting a 200-fold more powerful sweetness effect than conventional sugars (Magnuson, Carakostas, Moore, Poulos, & Renwick, 2016). Although the exact molecular mechanisms remain incompletely understood (DuBois, 2016), the consequence is that subjective sweetness ceases to track nutritional value. In other words, the pursuit of the proxy (whatever the receptor binds) has undermined its relation to the original evolutionary goal of that receptor (identifying energy-dense sugar molecules). Similar accounts emerge for some cases of neural or ecological proxy failure when they are traced down to the molecular level – they often involve "nonspecific" binding of receptors that play key roles in regulatory systems (see sects. 4.1 and 4.3).

The case of aspartame is particularly interesting, however, because it seems to reflect proxy *appropriation* rather than proxy failure: Proxy failure with respect to the goal of ingesting sugar would seem to imply undernourishment, but consumers choose sweeteners precisely because they allow sweet sensations without high calorific value. The fact that aspartame undermines the original purpose of both the molecular receptors and the neural reward system, thus becomes a feature. Our conscious self can hack our own reward system by intervening at the molecular level. Although the excessive use of artificial sweeteners entails problems of its own (Borges et al., 2017; Hsiao & Wang, 2013), this phenomenon nevertheless illustrates that lower-level proxy failure can be harnessed to promote higher-level goals, allowing us to overcome narrow evolutionary dictates.

Above, we have observed the interaction between proxies from the molecular level to the level of the conscious individual. Remarkably, the ramifications of the proxy nature of the T1 receptor can also be traced further up the levels to social and economic proxy systems. The choices of individuals to consume aspartame translate to the profitability of products containing aspartame. In other words, the molecular proxy performance determines the market-level proxy performance, directing human activities and resources towards the production, processing, distribution, marketing, and regulation of the artificial sweetener. In this example, the goals of conscious individuals are arguably what directs the whole system, harnessing proxy failure at lower levels and guiding proxies at higher levels. To what extent this can be generalized to other chemical components of our remarkably unexplored modern diets remains to be explored (Barabási, Menichetti, & Loscalzo, 2019). Indeed, corporate practices that harness addictive dynamics (e.g., in fast food; Small & DiFeliceantonio, 2019) appear to reflect additional instances, in which proxy failure at a lower level ("disrupted gut–brain signalling") is harnessed to serve higher-level goals (profitability; Box 8).

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Box 8. On "informative" vs. "persuasive" advertising (and marketing in general)

A perennial debate within economics concerns the question of whether corporate practices such as advertising primarily serve consumer or corporate goals (Akerlof & Shiller, 2015; Becker & Stigler, 1977; Hodgson, 2003; Susser, Roessler, & Nissenbaum, 2019). By the "informative" view, advertising is predominantly "informative" and not intended to change consumer "tastes" (Becker & Stigler, 1977). The upshot is that advertising strictly serves consumer goals (Kirkpatrick, 1994; Rizzo & Whitman, 2019).

By contrast, the "persuasive" view suggests that advertising (or marketing more generally; Franklin, Ashton, Gorman, & Armstrong, 2022) can *manipulate* consumer behaviour and preferences (Hodgson, 2003; Packard, 1958; Zuboff, 2019). This reflects an instance of *proxy appropriation* because consumers cease to be the "sovereign" of the market (Chirat, 2020; Galbraith, 1998). Instead, market-level proxies hack neural proxies to serve the higher-level goal of maximizing profitability and consumption (Braganza, 2022a, 2022b). Topical examples abound: For instance, "dark patterns" are digital interfaces designed to increase profits by deceiving or manipulating consumer (Luguri & Strahilevitz, 2021; Mathur et al., 2019). Descriptions of such practices in both digital (Zuboff, 2019) and nondigital (Akerlof & Shiller, 2015; Hanson & Kysar, 1999) contexts are too numerous to list. We subsume within this view any marketing practice that harnesses addictive consumer behaviour (Hadland, Rivera-Aguirre, Marshall, & Cerdá, 2019; Newall, 2019; Stuckler, McKee, Ebrahim, & Basu, 2012). Note the close relation of the present issue and the controversies discussed in Boxes 2 and 4.

Exploring these cases suggests two important observations. The first is that categorizing a phenomenon as proxy failure requires attributing a specific goal to the system. It is often natural to consider the goals of conscious individuals. But it is important to recognize that the system may ultimately be guided by goals at other levels. The second is that proxies at one level tend to be goals at other levels, raising the question to which degree a chain or web of proxies is sufficient to understand proxy failure, and teleonomic behaviour more generally. Goals appear as abstract objects, which help us understand why proxies are the way they are or how they might develop into the future. Yet the only tangible footprints that goals leave behind in the physical world are proxies.

6. Conclusion and outlook

> Certain forms of knowledge and control require a narrowing of vision. The great advantage of such tunnel vision is that it brings into sharp focus certain limited aspects of an otherwise far more complex and unwieldy reality. This very simplification, in turn, makes the phenomenon at the centre of the field of vision more legible and hence more susceptible to careful measurement and calculation. Combined with similar observations, an overall aggregate, synoptic view of a selective reality is achieved, making possible a high degree of schematic knowledge, control, and manipulation.
>
> *Seeing Like a State* (Scott, 2008)

In this paper we have argued that a plethora of diverse phenomena across biological and social scales can be categorized as *proxy failure*. Specifically, we suggest that whenever a *regulator* seeks to pursue a *goal* by incentivizing or selecting *agents* based on their production of a *proxy*, a pressure arises that tends to push the proxy away from the goal. We have explored examples of proxy failure in three domains, namely neuroscience, economics, and ecology, drawing attention to numerous similarities concerning both drivers and constraints of the phenomenon. We then outlined three characteristic consequences of proxy failure, which similarly recur across natural and social systems: The *proxy treadmill* suggests that signalling equilibria will often be unstable, tend to inflation, and drive increasing complexity. The *proxy cascade* suggests that, within nested hierarchical systems, higher levels can constrain proxy failure at lower levels, allowing goal-directed behaviour of the entire system. Finally, *proxy appropriation* suggests that higher-level proxies can *harness* proxy failure at lower levels to serve higher-level goals. The proxy perspective thus offers a powerful lens to explore how goal-oriented systems can achieve, or fail to achieve, their goals, and how the tension between both can drive complexity and coordination across biological and social systems.

Our account provides a theoretical unification across diverse disciplines and terminologies. However, the proposed synthesis is far from exhaustive: There are numerous instances of proxy failure in areas we have neglected (e.g., Muller, 2018). Further, over the course of writing this paper we have come to realize that *formal* theories and models of proxy failure tend to be highly domain specific: It is unclear how they relate to each other, or how a formal model of the unified mechanism presently outlined might look. Proxy failure draws attention to a variety of difficult conceptual questions, including about the appropriateness of using analytic equilibrium modelling to study what may be better understood as out-of-equilibrium phenomena. It also touches on a range of controversies within disciplines, such as (in biology) whether signals are generally honest or (in economics) whether revealed preferences are generally normative. We have briefly discussed some of these issues in Boxes 2–8. But there is clearly much need of further investigation and debate. For these reasons we view open peer commentary as particularly well-suited to the topic: We look forward to additional examples of proxy failure and fresh theoretical considerations.

The proxy failure framework also dovetails with important lessons from the chequered history of human social engineering. In *Seeing Like a State*, the anthropologist James C. Scott describes how attempts to impose rationalized, static regulatory models onto populations routinely fail. The reason, we have argued, is that exploratory behaviour aimed at proxy maximization will eventually uncover the blind spots of a regulatory model. Our account suggests that the risk of proxy failure should never be underestimated, particularly when proxies have become established. In the words of Bryar and Carr (2021) "unless you have a regular process to *independently validate the metric*, assume that over time something will cause it to drift and skew the numbers" (emphasis added). In the worst case, proxies continue to be observed uncritically and effectively "displace the goal" (Merton, 1940; Wouters, 2020). Muller (2018) observes that proxies often direct "time and resources away from doing and towards documenting" that can lead to a "mushroom-like growth of administrative staff." This may have the additional effect of "crowding out" intrinsic incentives (Frey & Jegen, 2001): As teachers or doctors are increasingly micromanaged, they will lose both the time and the will to perform aspects of their job that are difficult to measure. The result is a system with disillusioned practitioners, in which an ever-increasing share of total resources is directed towards increasing and measuring proxies (Muller, 2018).

But in the best case, proxy failure is continuously assessed and proxies are modified, deemphasized, or discarded as appropriate. There are numerous examples where this appears to have happened: Akerlof and Shiller (2015) describe a process of continuous hacking and counter-hacking between regulators and pharmaceutical companies over the course of the twentieth century. Despite many successful attempts by agents to subvert

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

regulation, it seems clear that the overall safety of pharmaceuticals has dramatically improved since the times when arsenic was sold as a universal cure. Similarly, dynamic regulation has contributed to increases in the safety of many consumer products and workplaces. Finally, in some cases it may be best to pause – and ask if the costs of an externally imposed proxy system outweigh its benefits (Muller, 2018). The present perspective suggests that, when we do opt for proxies, then one of the greatest risks to real progress will be to underestimate the risk of proxy failure.

Acknowledgements. We thank David Haig, the participants of the Bonn AI-ethics seminar, and six anonymous reviewers for helpful comments on the draft.

Financial support. Y. J. J. was supported by the US National Institutes of Mental Health Grant Nos. 01MH057414 and R01MH117785 (both awarded to Helen Barbas). D. E. M. was supported by Stanford Science Fellowship and the NSF Postdoctoral Research Fellowships in Biology PRFB Program, grant 2109465. O. B. was supported by grants from Bonfor (GEROK) and VW Foundation (Originalitätsverdacht).

Competing interest. None.

References

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. *The Journal of Political Economy*, *105*(1), 1–29.

Ågren, J. A., & Clark, A. G. (2018). Selfish genetic elements. *PLoS Genetics*, *14*(11), e1007700. https://doi.org/10.1371/JOURNAL.PGEN.1007700

Ågren, J. A., Davies, N. G., & Foster, K. R. (2019). Enforcement is central to the evolution of cooperation. *Nature Ecology & Evolution*, *3*(7), 1018–1029. https://doi.org/10.1038/s41559-019-0907-1

Ågren, J. A., & Patten, M. M. (2022). Genetic conflicts and the case for licensed anthropomorphizing. *Behavioral Ecology and Sociobiology*, *76*(12), 166. https://doi.org/10.1007/s00265-022-03267-6

Akerlof, G. A., & Shiller, R. J. (2015). *Phishing for phools: The economics of manipulation and deception* (pp. 1–288). Princeton University Press.

Albo, M. J., Winther, G., Tuni, C., Toft, S., & Bilde, T. (2011). Worthless donations: Male deception and female counter play in a nuptial gift-giving spider. *BMC Evolutionary Biology*, *11*, 329. https://doi.org/10.1186/1471-2148-11-329

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *ArXiv:1606.06565*. http://arxiv.org/abs/1606.06565

Anderson, B. A., Kuwabara, H., Wong, D. F., Gean, E. G., Rahmim, A., Brašić, J. R., … Yantis, S. (2016). The role of dopamine in value-based attentional orienting. *Current Biology: CB*, *26*(4), 550–555. https://doi.org/10.1016/j.cub.2015.12.062

Andersson, M., & Simmons, L. W. (2006). Sexual selection and mate choice. *Trends in Ecology & Evolution*, *21*(6), 296–302. https://doi.org/10.1016/J.TREE.2006.03.015

Arathi, H. S., Ganeshaiah, K. N., Shaanker, R. U., & Hegde, S. G. (1996). Factors affecting embryo abortion in *Syzygium cuminii* (L.) Skeels (Myrtaceae). *International Journal of Plant Sciences*, *157*(1), 49–52. https://doi.org/10.1086/297319

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, *58*(4), 328–346. https://doi.org/10.1086/256963

Ashton, H. (2020). *Causal Campbell–Goodhart's law and reinforcement learning. ArXiv:2011.01010*. http://arxiv.org/abs/2011.01010

Backwell, P. R. Y., Christy, J. H., Telford, S. R., Jennions, M. D., & Passmore, J. (2000). Dishonest signalling in a fiddler crab. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1444), 719–724. https://doi.org/10.1098/rspb.2000.1062

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *Journal of Human Resources*, *37*(4), 728–751.

Barabási, A. L., Menichetti, G., & Loscalzo, J. (2019). The unmapped chemical complexity of our diet. *Nature Food*, *1*(1), 33–37. https://doi.org/10.1038/s43016-019-0005-1

Bardoscia, M., Battiston, S., Caccioli, F., & Caldarelli, G. (2017). Pathways towards instability in financial networks. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/ncomms14416

Beale, N., Battey, H., Davison, A. C., & MacKay, R. S. (2020). An unethical optimization principle. *Royal Society Open Science*, *7*(7), 200462. https://doi.org/10.1098/rsos.200462

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*(10), 1154–1165. https://doi.org/10.1016/J.VISRES.2008.07.012

Becker, G., & Stigler, G. (1977). De gustibus non est disputandum. *American Economic Review*, *67*(2), 76–90.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*(4), 675–700.

Bénabou, R., & Tirole, J. (2016). Bonus culture: Competitive pay, screening, and multitasking. *Journal of Political Economy*, *124*(2), 305–370. https://doi.org/10.3386/w18936

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*(6), 787–793. https://doi.org/10.1038/s41593-018-0152-y

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *The American Psychologist*, *71*(8), 670–679. https://doi.org/10.1037/amp0000059

Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, *92*(8-9), 1787–1794.

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). Users polarization on Facebook and YouTube. *PLoS ONE*, *11*(8), e0159641. https://doi.org/10.1371/journal.pone.0159641

Biagioli, M., & Lippman, A. (2020). *Gaming the metrics: Misconduct and manipulation in academic research*. MIT Press. https://ieeexplore.ieee.org/book/9072252

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4–5), 303–345. https://doi.org/10.1016/S0361-3682(01)00052-6

Borges, M. C., Louzada, M. L., de Sá, T. H., Laverty, A. A., Parra, D. C., Garzillo, J. M. F., … Millett, C. (2017). Artificially sweetened beverages and the response to the global obesity crisis. *PLoS Medicine*, *14*(1), e1002195. https://doi.org/10.1371/journal.pmed.1002195

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bradshaw, S. (2019). Disinformation optimised: Gaming search engine algorithms to amplify junk news. *Internet Policy Review*, *8*(4), 1–24. https://doi.org/10.14763/2019.4.1442

Braganza, O. (2020). A simple model suggesting economically rational sample-size choice drives irreproducibility. *PLoS ONE*, *15*(3), e0229615. https://doi.org/10.1371/journal.pone.0229615

Braganza, O. (2022a). *Market paternalism – Do people really want to be nudged towards consumption?* IfSO Working Paper. https://www.uni-due.de/imperia/md/content/soziooekonomie/ifsowp23_braganza2022.pdf

Braganza, O. (2022b). Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, *9*(2), 1–41. https://doi.org/10.1098/RSOS.211030

Breslin, P. A. S. (2013). An evolutionary perspective on food and human taste. *Current Biology*, *23*(9), R409–R418. https://doi.org/10.1016/j.cub.2013.04.010

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. https://doi.org/10.1016/J.NEURON.2010.11.022

Bryar, C., & Carr, B. (2021). *Working backwards – Insights, stories and secrets from inside Amazon*. Macmillan USA.

Bullock, D., Tan, C. O., & John, Y. J. (2009). Computational perspectives on forebrain microcircuits implicated in reinforcement learning, action selection, and cognitive control. *Neural Networks*, *22*(5–6), 757–765. https://doi.org/10.1016/j.neunet.2009.06.008

Burnham, T. C. (2016). Economics and evolutionary mismatch: Humans in novel settings do not maximize. *Journal of Bioeconomics*, *18*(3), 195–209. https://doi.org/10.1007/s10818-016-9233-8

Buss, L. W. (1988). *The evolution of individuality*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691632858/the-evolution-of-individuality

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Casarini, L., Santi, D., Brigante, G., & Simoni, M. (2018). Two hormones for one receptor: Evolution, biochemistry, actions, and pathophysiology of LH and hCG. *Endocrine Reviews*, *39*(5), 549–592. https://doi.org/10.1210/er.2018-00065

Chirat, A. (2020). A reappraisal of Galbraith's challenge to consumer sovereignty: Preferences, welfare and the non-neutrality thesis. *European Journal of the History of Economic Thought*, *27*(2), 248–275. https://doi.org/10.1080/09672567.2020.1720763

Coddington, L. T., & Dudman, J. T. (2019). Learning from action: Reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, *104*(1), 63–77. https://doi.org/10.1016/J.NEURON.2019.08.036

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39–67. https://doi.org/10.1177/0149206310388419

Cowen, N., & Dold, M. F. (2021). Introduction: Symposium on escaping paternalism: Rationality, behavioral economics and public policy by Mario J. Rizzo and Glen Whitman. *Review of Behavioral Economics*, *8*(3-4), 213–220.

Cruz, B. F., & Paton, J. J. (2021). Dopamine gives credit where credit is due. *Neuron*, *109*(12), 1915–1917. https://doi.org/10.1016/J.NEURON.2021.05.033

Csiszar, A. (2020). Gaming metrics before the game: Citation and the bureaucratic virtuoso. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 31–42). MIT Press. https://ieeexplore.ieee.org/document/9085771

Dawkins, M. S., & Guilford, T. (1991). The corruption of honest signalling. *Animal Behaviour*, *41*(5), 865–873. https://doi.org/10.1016/S0003-3472(05)80353-7

Demski, A., & Garrabrant, S. (2020). Embedded agency. *ArXiv*.

Dennett, D. (2009). Intentional systems theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *Oxford handbook of philosophy of mind* (pp. 339–350). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., … Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1595–1600. https://doi.org/10.1073/pnas.1417219112

DeYoung, C. G. (2013). The neuromodulator of exploration: A unifying theory of the role of dopamine in personality. *Frontiers in Human Neuroscience*, *7*, 1–26. https://doi.org/10.3389/fnhum.2013.00762

Diana, M. (2011). The dopamine hypothesis of drug addiction and its potential therapeutic value. *Frontiers in Psychiatry*, *2*, 1–7. https://www.frontiersin.org/articles/10.3389/fpsyt.2011.00064

Dold, M. F., & Schubert, C. (2018). Toward a behavioral foundation of normative economics. *Review of Behavioral Economics*, *5*(3–4), 221–241. https://doi.org/10.1561/105.00000097

DuBois, G. E. (2016). Molecular mechanism of sweetness sensation. *Physiology & Behavior*, *164*, 453–463. https://doi.org/10.1016/j.physbeh.2016.03.015

Dyson, E. A., & Hurst, G. D. D. (2004). Persistence of an extreme sex-ratio bias in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6520–6523. https://doi.org/10.1073/pnas.0304068101

Edmans, A., Fang, V. W., & Lewellen, K. A. (2017). Equity vesting and investment. *The Review of Financial Studies*, *30*(7), 2229–2271. https://doi.org/10.1093/rfs/hhx018

Ellis, G. F. R., & Kopel, J. (2019). The dynamical emergence of biology from physics: Branching causation via biomolecules. *Frontiers in Physiology*, *9*, 1966. https://doi.org/10.3389/fphys.2018.01966

Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, *198*, 6435–6467. https://doi.org/10.1007/S11229-021-03141-4

Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy videos. *ArXiv*. http://arxiv.org/abs/2003.03318

Farnsworth, K. D., Albantakis, L., & Caruso, T. (2017). Unifying concepts of biological function from molecules to ecosystems. *Oikos*, *126*(10), 1367–1376. https://doi.org/10.1111/oik.04171

Fellner, W. J., & Goehmann, B. (2020). Human needs, consumerism and welfare. *Cambridge Journal of Economics*, *44*(2), 303–318. https://doi.org/10.1093/cje/bez046

Finan, F., & Schechter, L. (2012). Vote-buying and reciprocity. *Econometrica*, *80*(2), 863–881. https://doi.org/10.3982/ECTA9035

Finefter-Rosenbluh, I., & Levinson, M. (2015). Philosophical inquiry in education. *The Journal of the Canadian Philosophy of Education Society*, *23*, 3–21. https://journals.sfu.ca/pie/index.php/pie/article/view/894

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's law in action. *GigaScience*, *8*(6), 1–20. https://doi.org/10.1093/gigascience/giz053

Fisher, R. (1930). *The genetical theory of natural selection*. Clarendon Press. https://books.google.com/books?hl=de&lr=&id=WPfvAgAAQBAJ&oi=fnd&pg=PA1&ots=onzSFD_Yvp&sig=r7RthPTyPdIjR2p4PCpeIHvxnNg

Flynn, T. T. (1922). The phylogenetic significance of the marsupial allantoplacenta.

Franco-Santos, M., & Otley, D. (2018). Reviewing and theorizing the unintended consequences of performance management systems. *International Journal of Management Reviews*, *20*(3), 696–730. https://doi.org/10.1111/ijmr.12183

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good*. Princeton University Press.

Franken, I. H. A., Booij, J., & van den Brink, W. (2005). The role of dopamine in human addiction: From reward to motivated attention. *European Journal of Pharmacology*, *526*(1–3), 199–206. https://doi.org/10.1016/J.EJPHAR.2005.09.025

Franklin, M., Ashton, H., Gorman, R., & Armstrong, S. (2022). *Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI*.

Fremstad, A., & Paul, M. (2022). Neoliberalism and climate change: How the free-market myth has prevented climate action. *Ecological Economics*, *197*, 107353. https://doi.org/10.1016/J.ECOLECON.2022.107353

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611. https://doi.org/10.1111/1467-6419.00150

Friedman, E. J., & Oren, S. S. (1995). The complexity of resource allocation and price mechanisms under bounded rationality. *Economic Theory*, *6*(2), 225–250. https://doi.org/10.1007/BF01212489

Frutiger, A., Tanno, A., Hwu, S., Tiefenauer, R. F., Vörös, J., & Nakatsuka, N. (2021). Nonspecific binding – Fundamental concepts and consequences for biosensing applications. *Chemical Reviews*, *121*(13), 8095–8160. https://doi.org/10.1021/acs.chemrev.1c00044

Funk, D. H., & Tallamy, D. W. (2000). Courtship role reversal and deceptive signals in the long-tailed dance fly, *Rhamphomyia longicauda*. *Animal Behaviour*, *59*(2), 411–421. https://doi.org/10.1006/anbe.1999.1310

Galbraith, J. K. (1998). *The affluent society*. Houghton Mifflin.

Gasparini, C., Serena, G., & Pilastro, A. (2013). Do unattractive friends make you look better? Context-dependent male mating preferences in the guppy. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1756), 20123072. https://doi.org/10.1098/rspb.2012.3072

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108* (Suppl. 3), 15647–15654. https://doi.org/10.1073/pnas.1014269108

Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, *1*, 1–20. https://doi.org/10.1007/978-1-349-17295-5_4

Gray, M. W., Lukeš, J., Archibald, J. M., Keeling, P. J., & Doolittle, W. F. (2010). Irremediable complexity? *Science*, *330*(6006), 920–921. https://doi.org/10.1126/science.1198594

Grether, G. F., Hudon, J., & Endler, J. A. (2001). Carotenoid scarcity, synthetic pteridine pigments and the evolution of sexual coloration in guppies (*Poecilia reticulata*). *Proceedings of the Royal Society B: Biological Sciences*, *268*(1473), 1245. https://doi.org/10.1098/RSPB.2001.1624

Griesemer, J. (2020). Taking Goodhart's law meta: Gaming, meta-gaming, and hacking academic performance metrics. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 77–87). MIT Press.

Hadland, S. E., Rivera-Aguirre, A., Marshall, B. D. L., & Cerdá, M. (2019). Association of pharmaceutical industry marketing of opioid products with mortality from opioid-related overdoses. *JAMA Network Open*, *2*(1), e186007. https://doi.org/10.1001/jamanetworkopen.2018.6007

Haig, D. (1993). Genetic conflicts in human pregnancy. *The Quarterly Review of Biology*, *68*(4), 495–532. https://doi.org/10.1086/418300

Haig, D. (2020). *From Darwin to Derrida: Selfish genes, social selves, and the meanings of life*. MIT Press.

Hanson, J., & Kysar, D. A. (1999). Taking behavioralism seriously: The problem of market manipulation. *New York University Law Review*, *74*, 632.

Hardin, R., & Cullity, G. (2020). The free rider problem. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (pp. 1–20). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/free-rider/

Harris, K. D., Daon, Y., & Nanjundiah, V. (2020). The role of signaling constraints in defining optimal marginal costs of reliable signals. *Behavioral Ecology*, *31*(3), 784–791. https://doi.org/10.1093/beheco/araa025

Hartman, C. G. (1920). Studies in the development of the opossum *Didelphys virginiana* L. V. The phenomena of parturition. *The Anatomical Record*, *19*(5), 251–261. https://doi.org/10.1002/ar.1090190502

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, *402*(S6761), C47–C52. https://doi.org/10.1038/35011540

Heather, N. (2017). Q: Is addiction a brain disease or a moral failing? A: Neither. *Neuroethics*, *10*(1), 115–124. https://doi.org/10.1007/s12152-016-9289-0

Henke, A., & Gromoll, J. (2008). New insights into the evolution of chorionic gonadotrophin. *Molecular and Cellular Endocrinology*, *291*(1), 11–19. https://doi.org/10.1016/j.mce.2008.05.009

Hennessy, C., & Goodhart, C. A. E. (2021). Goodhart's law and machine learning. SSRN 3639508. https://doi.org/10.2139/ssrn.3639508

Heylighen, F., & Joslyn, C. (2001). Cybernetics and second-order cybernetics. In R. A. Meyers (Ed.), *Encyclopedia of physical science & technology* (pp. 1–24). Academic Press.

Hill, J. P., & Hill, W. C. O. (1955). The growth-stages of the pouch-young of the native cat (*Dasyurus viverrinus*) together with observations on the anatomy of the new-born young. *The Transactions of the Zoological Society of London*, *28*(5), 349–352. https://doi.org/10.1111/j.1096-3642.1955.tb00003.x

Hodgson, G. M. (2003). The hidden persuaders: Institutions and individuals in economic theory. *Cambridge Journal of Economics*, *27*(2), 159–175. https://doi.org/10.1093/CJE/27.2.159

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, *10* (1), 74. ttps://doi.org/10.2307/3003320

Holmström, B. (2017). Pay for performance and beyond. *American Economic Review*, *107* (7), 1753–1777. https://doi.org/10.1257/aer.107.7.1753

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17* (66), 159. https://doi.org/10.2307/2549382

Hsiao, A., & Wang, Y. C. (2013). Reducing sugar-sweetened beverage consumption: Evidence, policies, and economics. *Current Obesity Reports*, *2*(3), 191–199. https://doi.org/10.1007/s13679-013-0065-8

Ittner, C. D., Larcker, D. F., & Meyer, M. W. (1997). *Performance, compensation, and the balanced scorecard*. Wharton School, University of Pennsylvania.

Iwasa, Y., & Pomiankowski, A. (1995). Continual change in mate preferences. *Nature*, *377* (6548), 420–422. https://doi.org/10.1038/377420a0

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Jeckelmann, J.-M., & Erni, B. (2020). Transporters of glucose and other carbohydrates in bacteria. *Pflugers Archiv: European Journal of Physiology*, *472*(9), 1129–1153. https://doi.org/10.1007/s00424-020-02379-0

Jones, B. A., & Rachlin, H. (2009). Delay, probability, and social discounting in a public goods game. *Journal of the Experimental Analysis of Behavior*, *91*(1), 61–73. https://doi.org/10.1901/jeab.2009.91-61

Kauffman, S. A. (2019). *A world beyond physics: The emergence and evolution of life*. Oxford University Press. https://scholar.google.com/scholar_lookup?title=A+World+beyond+Physics:+The+Emergence+and+Evolution+of+Life&author=Stuart,+K.&publication_year=2019

Kelly, C., & Snower, D. J. (2021). Capitalism recoupled. *Oxford Review of Economic Policy*, *37*(4), 851–863. https://doi.org/10.1093/oxrep/grab025

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, *18*(4), 769–783. https://doi.org/10.5465/255378

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

Kirkpatrick, J. (1994). *Defense of advertising: Arguments from reason, ethical egoism, and laissez-faire capitalism*. https://philpapers.org/rec/KIRIDO

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *Journal of Neuroscience*, *28*(31), 7837–7846. https://doi.org/10.1523/JNEUROSCI.1600-08.2008

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Krebs, J. R., & Dawkins, R. (1984). Animal Signals: Mind-Reading and Manipulation. In J. R. Krebs, & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (S. 380–402). Blackwell Scientific.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. https://doi.org/10.1016/J.NEURON.2009.09.040

Ledford, J. L. (2016). *Search engine optimization Bible*. Wiley. https://books.google.de/books?hl=de&lr=&id=2Gz-CAAAQBAJ&oi=fnd&pg=PR16&dq=search+engine+optimization&ots=vIGJQrdX1C&sig=h6Lx7esR7Tad8aBjyrOY8SPNARQ#v=onepage&q=search engine optimization&f=false

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/J.CONB.2012.06.001

Lewis, D. A., & Sesack, S. R. (1997). Chapter VI – Dopamine systems in the primate brain. In F. E. Bloom, A. Björklund, & T. Hökfelt (Eds.), *Handbook of chemical neuroanatomy* (Vol. 13, pp. 263–375). Elsevier. https://doi.org/10.1016/S0924-8196(97)80008-5

Leyton, M., & Vezina, P. (2014). Dopamine ups and downs in vulnerability to addictions: A neurodevelopmental model. *Trends in Pharmacological Sciences*, *35*(6), 268–276. https://doi.org/10.1016/j.tips.2014.04.002

Li, X., Staszewski, L., Xu, H., Durick, K., Zoller, M., & Adler, E. (2002). Human receptors for sweet and umami taste. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(7), 4692–4696. https://doi.org/10.1073/pnas.072090199

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, *1*, 19–46. https://doi.org/10.1016/S0167-2231(76)80003-6

Luguri, J., & Strahilevitz, L. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, *13*(1), 43–109. https://doi.org/10.2139/ssrn.3431205

Lüpold, S., Manier, M. K., Puniamoorthy, N., Schoff, C., Starmer, W. T., Luepold, S. H. B., … Pitnick, S. (2016). How sexual selection can drive the evolution of costly sperm ornamentation. *Nature*, *533*(7604), 535–538. https://doi.org/10.1038/nature18005

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(Suppl. 1), 8597–8604. https://doi.org/10.1073/pnas.0702207104

Magnuson, B. A., Carakostas, M. C., Moore, N. H., Poulos, S. P., & Renwick, A. G. (2016). Biological fate of low-calorie sweeteners. *Nutrition Reviews*, *74*(11), 670–689. https://doi.org/10.1093/nutrit/nuw032

Manheim, D. (2018). *Overoptimization failures and specification gaming in multi-agent systems. ArXiv:1810.10862*. https://arxiv.org/abs/1810.10862v2

Manheim, D., & Garrabrant, S. (2018). *Categorizing variants of Goodhart's law. ArXiv:1803.04585v4*. https://arxiv.org/abs/1803.04585v3

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory* (Vol. 1). Oxford University Press.

Masur, J. S., & Posner, E. A. (2015). Toward a Pigouvian state. *University of Pennsylvania Law Review*, *164*(1), 93–147.

Mathur, A., Friedman, M. J., Mayer, J., Narayanan, A., Acar, G., Lucherini, E., & Chetty, M. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–32. https://doi.org/10.1145/3359183

Mayer, C. (2021). The future of the corporation and the economics of purpose. *Journal of Management Studies*, *58*(3), 887–901. https://doi.org/10.1111/joms.12660

Maynard-Smith, J., & Harper, D. (2003). *Animal signals*. Oxford University Press. https://books.google.de/books?hl=de&lr=&id=SUA51MeG1lcC&oi=fnd&pg=PA1&dq=animal + signals&ots=zOIsmDUlw6&sig=Gw_lUql1LU-HHVU3cfNuJiOG924

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*(3489), 1501–1506.

McCoy, D. E., & Haig, D. (2020). Embryo selection and mate choice: Can "honest signals" be trusted? *Trends in Ecology & Evolution*, *35*(4), 308–318. https://doi.org/10.1016/J.TREE.2019.12.002

McCoy, D. E., Shultz, A. J., Vidoudez, C., van der Heide, E., Dall, J. E., Trauger, S. A., & Haig, D. (2021). Microstructures amplify carotenoid plumage signals in tanagers. *Scientific Reports*, *11*(1), 1–20. https://doi.org/10.1038/s41598-021-88106-w

McElroy, J. S. (2014). Vavilovian mimicry: Nikolai Vavilov and his little-known impact on weed science. *Weed Science*, *62*(2), 207–216. https://doi.org/10.1614/WS-D-13-00122.1

McShea, D. W. (2016). Freedom and purpose in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *58*, 64–72. https://doi.org/10.1016/J.SHPSC.2015.12.002

Merton, R. K. (1940). Bureaucratic structure and personality. *Social Forces*, *18*(4), 560–568. https://doi.org/10.2307/2570634

Mink, J. W. (2018). Basal ganglia mechanisms in action selection, plasticity, and dystonia. *European Journal of Paediatric Neurology: EJPN*, *22*(2), 225–229. https://doi.org/10.1016/j.ejpn.2018.01.005

Mirowski, P., & Nik-Khah, E. (2017). *The knowledge we have lost in information* (Vol. 1). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190270056.001.0001

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Newall, P. W. S. (2019). Dark nudges in gambling. *Addiction Research & Theory*, *27*(2), 65–67. https://doi.org/10.1080/16066359.2018.1474206

Nichols, S. L., & Berliner, D. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. The Great Lakes Center for Education Research & Practice.

Okasha, S. (2006). *Evolution and the levels of selection*. Clarendon Press. https://books.google.de/books?hl=de&lr=&id=aw9REAAAQBAJ&oi=fnd&pg=PR9&dq=Evolution+and+the+levels+of+selection&ots=Dm690QpPDd&sig=NVz-SN_3wN_i_DpEfr-fjgWdIKo#v=onepage&q=Evolution and the levels of selection&f=false

O'Mahony, S. (2018). Medicine and the McNamara fallacy. *Journal of the Royal College of Physicians of Edinburgh*, *47*(3), 281–287. https://doi.org/10.4997/JRCPE.2017.315

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Packard, V. (1958). *The hidden persuaders*. Pocket Books.

Pearce, J. A. (1982). Problems facing first-time managers. *Human Resource Management*, *21*(1), 35–38. https://doi.org/10.1002/hrm.3930210108

Penn, D. J., & Számadó, S. (2020). The handicap principle: How an erroneous hypothesis became a scientific principle. *Biological Reviews*, *95*(1), 267–290. https://doi.org/10.1111/brv.12563

Perez-Aguilar, J. M., Kang, S., Zhang, L., & Zhou, R. (2019). Modeling and structural characterization of the sweet taste receptor heterodimer. *ACS Chemical Neuroscience*, *10*(11), 4579–4592. https://doi.org/10.1021/acschemneuro.9b00438

Peters, A., Schweiger, U., Pellerin, L., Hubold, C., Oltmanns, K. M., Conrad, M., … Fehm, H. L. (2004). The selfish brain: Competition for energy resources. *Neuroscience & Biobehavioral Reviews*, *28*(2), 143–180. https://doi.org/10.1016/J.NEUBIOREV.2004.03.002

Poku, M. (2016). Campbell's law: Implications for health care. *Journal of Health Services Research & Policy*, *21*(2), 137–139. https://doi.org/10.1177/1355819615593772

Price, I., & Clark, E. (2009). An output approach to property portfolio performance measurement. *Property Management*, *27*(1), 6–15.

Prum, R. O. (2010). The Lande–Kirkpatrick mechanism is the null model of evolution by intersexual selection: Implications for meaning, honesty, and design in intersexual signals. *Evolution*, *64*, 3085–3100.

Prum, R. O. (2012). Aesthetic evolution by mate choice: Darwin's really dangerous idea. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1600), 2253–2265. https://doi.org/10.1098/RSTB.2011.0285

Prum, R. O. (2017). *The evolution of beauty: How Darwin's forgotten theory of mate choice shapes the animal world*. Anchor.

Pueyo, S. (2018). Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, *197*, 1731–1736. https://doi.org/10.1016/J.JCLEPRO.2016.12.138

Puig, M. V., Rose, J., Schmidt, R., & Freund, N. (2014). Dopamine modulation of learning and memory in the prefrontal cortex: Insights from studies in primates, rodents, and birds. *Frontiers in Neural Circuits*, *8*(Aug), 93. https://doi.org/10.3389/FNCIR.2014.00093/BIBTEX

Raworth, K. (2018). *Doughnut economics: Seven ways to think like a 21st-century economist*. Random House Business.

Rendell, L., Fogarty, L., & Laland, K. N. (2011). Runaway cultural niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1566), 823–835. https://doi.org/10.1098/rstb.2010.0256

Reynolds, J. N. J., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70. https://doi.org/10.1038/35092560

Rizzo, M. J., & Whitman, G. (2019). *Escaping paternalism: Rationality, behavioral economics, and public policy* (pp. 1–496). Cambridge University Press. https://doi.org/10.1017/9781139061810

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

Roux, E. (2014). The concept of function in modern physiology. *The Journal of Physiology*, *592*(11), 2245–2249. https://doi.org/10.1113/jphysiol.2014.272062

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61. https://doi.org/10.2307/2548836

Satel, S., & Lilienfeld, S. O. (2014). Addiction and the brain-disease fallacy. *Frontiers in Psychiatry*, *4*, 141. https://doi.org/10.3389/fpsyt.2013.00141

Schultz, W. (2022). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32. https://doi.org/10.31887/DCNS.2016.18.1/WSCHULTZ

Scott, J. C. (2008). *Seeing like a state*. Yale University Press. https://doi.org/10.12987/9780300128789/HTML

Seo, M., Lee, E., & Averbeck, B. B. (2012). Action selection and action value in frontal-striatal circuits. *Neuron*, *74*(5), 947–960. https://doi.org/10.1016/j.neuron.2012.03.037

Shaanker, R. U., Ganeshaiah, K. N., & Bawa, K. S. (1988). Parent–offspring conflict, sibling rivalry, and brood size patterns in plants. *Annual Review of Ecology and Systematics*, *19*(1), 177–205. https://doi.org/10.1146/annurev.es.19.110188.001141

Shaffer, H. G. (1963). A new incentive for soviet managers. *The Russian Review*, *22*(4), 410–416. https://doi.org/10.2307/126674

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848–851. https://doi.org/10.1126/SCIENCE.1160575

Siebert, H. (2001). Der Kobra-Effekt wie man Irrwege der Wirtschaftspolitik vermeidet. Deutsche V.-A., Stgt.

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*(4507), 501–504. https://doi.org/10.1126/science.7244649

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Small, D. M., & DiFeliceantonio, A. G. (2019). Processed foods and food reward. *Science*, *363*(6425), 346–347. https://doi.org/10.1126/science.aav0556

Smith, J. E., & Winkler, R. L. (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, *52*(3), 311–322. https://doi.org/10.1287/mnsc.1050.0451

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355. https://doi.org/10.2307/1882010

Strathern, M. (1997). "Improving ratings": Audit in the British university system. *European Review*, *55*(5), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:33.0.CO;2-4

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, *11*(6), 800–816. https://doi.org/10.1177/1745691616650284

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112

Stuckler, D., McKee, M., Ebrahim, S., & Basu, S. (2012). Manufacturing epidemics: The role of global producers in increased consumption of unhealthy commodities including processed foods, alcohol, and tobacco. *PLoS Medicine*, *9*(6), e1001235. https://doi.org/10.1371/journal.pmed.1001235

Sugden, R. (2017). Do people really want to be nudged towards healthy lifestyles? *International Review of Economics*, *64*(2), 113–123.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, *4*, 1–45. https://doi.org/10.2139/ssrn.3306006

Számadó, S., & Penn, D. J. (2018). Does the handicap principle explain the evolution of dimorphic ornaments? *Animal Behaviour*, *138*, e7–e10. https://doi.org/10.1016/j.anbehav.2018.01.005

Tardieu, H., Daly, D., Esteban-Lauzán, J., Hall, J., & Miller, G. (2020). Measuring the transformation – KPIs for understanding transformation progress. In *Deliberately digital* (p. 202). Springer. https://doi.org/10.1007/978-3-030-37955-1_4

Tayan, B. (2019). The Wells Fargo cross-selling scandal. In *Rock Center for Corporate Governance at Stanford University Closer Look Series: Topics, Issues and Controversies in Corporate Governance No. CGRP-62 Version 2*. https://corpgov.law.harvard.edu/2019/02/06/the-wells-fargo-cross-selling-scandal-2/

Thaler, R. H. (2018). From cashews to nudges: The evolution of behavioral economics. *American Economic Review*, *108*(6), 1265–1287. https://doi.org/10.1257/aer.108.6.1265

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin Books.

The Lincoln Electric Company. (1975).

Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. *Patterns*, *3*(5), 100476. https://doi.org/10.1016/J.PATTER.2022.100476

Thomson, R., Royed, T., Naurin, E., Artés, J., Costello, R., Ennser-Jedenastik, L., … Praprotnik, K. (2017). The fulfillment of parties' election pledges: A comparative study on the impact of power sharing. *American Journal of Political Science*, *61*(3), 527–542. https://doi.org/10.1111/ajps.12313

van der Kolk, B. (2022). Numbers speak for themselves, or do they? On performance measurement and its implications. *Business & Society*, *61*(4), 813–817. https://doi.org/10.1177/00076503211068433

Vann, M. G. (2003). Of rats, rice, and race: The great Hanoi rat massacre, an episode in French colonial history. *French Colonial History*, *4*(1), 191–203. https://doi.org/10.1353/fch.2003.0027

Vavilov, N. I. (1922). The law of homologous series in variation. *Journal of Genetics*, *12*(1), 47–89. https://doi.org/10.1007/BF02983073

Volkow, N. D., Wise, R. A., & Baler, R. (2017). The dopamine motive system: Implications for drug and food addiction. *Nature Reviews Neuroscience*, *18*(12), 741–752. https://doi.org/10.1038/nrn.2017.130

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press. https://doi.org/10.2307/2019327

Weaver, R. J., Santos, E. S. A., Tucker, A. M., Wilson, A. E., & Hill, G. E. (2018). Carotenoid metabolism strengthens the link between feather coloration and individual quality. *Nature Communications*, *9*(1), 73. https://doi.org/10.1038/s41467-017-02649-z

Wickler, W. (1965). Mimicry and the evolution of animal communication. *Nature*, *208*(5010), 519–521. https://doi.org/10.1038/208519a0

Wiedmann, T., Lenzen, M., Keyßer, L. T., & Steinberger, J. K. (2020). Scientists' warning on affluence. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-16941-y

Willson, M. F., & Burley, N. (1983). *Mate choice in plants (MPB-19), volume 19: Tactics, mechanisms, and consequences. (MPB-19)*. Princeton University Press. https://doi.org/10.2307/j.ctvx5wbss

Wilson, D. S., & Kirman, A. P. (2016). *Complexity and evolution: Toward a new synthesis for economics*. MIT Press.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483–494. https://doi.org/10.1038/nrn1406

Wise, R. A., & Robble, M. A. (2020). Dopamine and addiction. *Annual Review of Psychology*, *71*, 79–106. https://doi.org/10.1146/ANNUREV-PSYCH-010418-103337

Wouters, P. (2020). The mismeasurement of quality and impact. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 67–76). MIT Press.

Yankelovich, D. (1972). *Corporate priorities: A continuing study of the new demands on business*. Yankelovich Inc.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476. https://doi.org/10.1038/nrn1919

Zahavi, A. (1975). Mate selection – A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Zahavi, A., Butlin, R. K., Guilford, T., & Krebs, J. R. (1993). The fallacy of conventional signalling. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *340*(1292), 227–230. https://doi.org/10.1098/rstb.1993.0061

Zeleznik, A. J. (1998). In vivo responses of the primate corpus luteum to luteinizing hormone and chorionic gonadotropin. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(18), 11002–11007. https://doi.org/10.1073/pnas.95.18.11002

Zhuang, S., & Hadfield-Menell, D. (2021). *Consequences of misaligned AI* (arXiv:2102.03896). arXiv. http://arxiv.org/abs/2102.03896

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

Appendix 1. A note on anthropomorphic language

We employ anthropomorphic terminology to describe both human and nonhuman (or nonconscious) decision-making systems simply because it is useful and concise (Ågren & Patten, 2022; Haig, 2020). Anthropomorphic terms – such as "goal," "decision," or "hack" – are often metaphorical, but avoiding them makes the discussion of proxy failure unnecessarily arduous. Our generalization rests on two concepts: *Teleonomy* and *selection by consequences*, which share broad currency (McCoy & Haig, 2020; Smaldino & McElreath, 2016).

Teleonomy (Mayr, 1961) or the "intentional stance" (Dennett, 2009) justifies the conceptualization of systems "as if" they had goals (Mayr, 1961; McShea, 2016) because it is useful and efficient, even if these goals are not explicitly or consciously represented anywhere. Teleonomy is applied widely in biology (Ellis & Kopel, 2019; Farnsworth et al., 2017; Hartwell, Hopfield, Leibler, & Murray, 1999; Roux, 2014) and underlies our use of anthropomorphic language in nonhuman contexts independent of debates about the ontological status of "goals."

Selection by consequences (Skinner, 1981) points out a functional equivalence between natural selection and behaviour. Behaviour can be seen as the selection of actions; just as natural selection reflects the selection of traits. In

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

both cases, it is the "consequences" of a given behaviour/trait that determine whether it will be selected. The implication is that reinforcement, competition, and selection can all be viewed as analogous optimization processes that influence the agents' actions or traits, amplifying some and inhibiting others. Note that by drawing on this specific concept, we by no means endorse Skinnerian views (e.g., behaviourism) more broadly.

Appendix 2. Relation to Manheim and Garrabrant (2018)

Manheim and Garrabrant (2018) suggest a categorization of (at least) four mechanisms by which the optimization of a proxy measure can lead to its decorrelation from the goal: (1) regressional, (2) extremal, (3) causal, and (4) adversarial Goodhart (see also Demski & Garrabrant, 2020, for intuitive visualizations). We will briefly introduce each mechanism and then outline how they are related to our framework.

(1) *Regressional Goodhart*, also known as the "Optimizer's Curse" (Smith & Winkler, 2006), is based on the observation that for any "noisy" (i.e., imperfect) correlation, the highest (i.e., optimal) values of one variable (the proxy) will tend to exceed the regression line, leading to a systematic "shortfall" with respect to the other variable (the goal) in the region where the proxy is optimized.
(2) *Extremal Goodhart* describes a situation in which "optimization pushes you outside the range where the correlation exists, into portions of the distribution which behave very differently" (Demski & Garrabrant, 2020). It is arguably a generalized version of regressional Goodhart for nonlinear relations.
(3) *Causal Goodhart* describes a situation in which an action to optimize the proxy destroys the initial correlation altogether. The main difference between (1) and (2) is that here a causal intervention in the real world is assumed, whereas the two previous cases pertain to purely inferential phenomena.
(4) *Adversarial Goodhart* describes a situation in which the use of the proxy induces "agents" to causally intervene in the real world, destroying the correlation that motivated the regulator to choose the proxy.

Although, Manheim and Garrabrant describe only the third case as causal, we think this is somewhat misleading, or rather a consequence of framing some cases as purely "inferential" rather than "control" problems:

Inferential Goodhart: Cases 1 and 2 reflect *inference* problems. No causal effects are assumed to happen in the "real world." The underlying distributions and causal networks linking proxy and goal are assumed as given and not modifiable (because an inferential procedure generally will not contain actuators). In these cases "when a measure becomes a target" means that the inferential algorithm, over the course of optimization, *moves* to a region of the data where the approximation becomes worse or breaks down. In this case, causation happens entirely within the inferential optimization procedure (the *movement*), such that it may not be seen as "real" causation. The inferential framing abstracts away from the causes that underlie the *noise* in "regressional" or the *nonlinearity* in "extremal" Goodhart (these causes would correspond to the causes that asymmetrically affect proxy and goal in our framework).

Control Goodhart: Cases 3 and 4 reflect *control* problems where a mechanism that causally affects the "real world" is assumed. The "measure becoming the target" changes the distributions of proxy and/or goal. For case 3, the optimization algorithm itself manipulates the proxy values. For case 4, the "incentivization" of some "adversary" leads these to cause changed distributions. Case 4 thus seems to be a special case of 3, where the regulator and agent are more easily associated with distinct goals.

Cases 1–3 correspond to cases where agents are simply passive entities, strategies, or actions, being selected by the regulator because they are associated with larger proxy values (cases 1 and 2) or cause the proxy to increase (case 3). Although we would argue that strict inference problems that lead to no action in the "real world" are of minor interest here, they may explain actions based on faulty inference. If we assume an action based on the inferences in cases 1 and 2, then the proxy failure becomes overtly causal. Consider a robot subject to inferential Goodhart: Once the robot acts on the faulty inference, he causes the measure to become a worse measure. Finally, cases 3 and 4 are functionally equivalent, because they simply differ in the mechanism by which the proxy is optimized (see Appendix 1). In this manner, all four cases can be related to our general mechanism.

Open Peer Commentary

Behavioral proxies compete by the time courses of their rewards, including endogenous rewards

George Ainslie

Department of Veterans Affairs Medical Center, Coatesville, PA, USA
www.picoeconomics.org

Corresponding author:
George Ainslie;
Email: info@picoeconomics.org

doi:10.1017/S0140525X23002960, e68

Abstract

Natural selection is slow, so behavioral goals must be based on patterns of reward. Addictions are rewarded in the same way as adaptive choice, so they can be distinguished only by their time course. In addition, the reward process is more plastic than is generally recognized, so abstract goals are shaped by the "legibility" of their proxies.

The authors' distinction between goal and proxy is especially useful in "preference learning" (target article, Table 1, sect. 4.1), where our understanding of selection by reward is in flux: (1) Given the necessarily slow pace of natural selection, the effective basis of behavioral goals must be reward. (2) "Addiction and other maladaptive habits" (target article, sect. 4.1, para. 1) depend on the same neural "behavior prioritization" (target article, sect. 4.1, para. 4) as adaptive choice, and can be distinguished only by their time course. (3) The reward process is more plastic than is generally recognized, so abstract goals are shaped by the "legibility" of their proxies.

(1) The authors depict the brain's goal as "the organism's fitness, utility, or wellbeing" (target article, sect. 4.1, para. 2) – evolutionary adaptiveness. But adaptiveness inevitably diverges from rewardingness. The action-selecting *reward mechanism* has clearly been shaped by natural selection, but the *rewards* it specifies become outdated when environmental circumstances change, as described in Box 2 in the target article. In the case of drug addictions, evolution is probably "trying" to correct the readiness of the reward mechanism to be hijacked by dopaminergic proxies – for instance, when it selects a gene that reduces the attraction of alcohol in some populations (Agarwal & Goedde, 1989) – but any such correction will take generations to occur. In the meantime, humans perceive welfare as a pattern of reward, and its connection with adaptiveness is purely historic. The connection may even be adversarial, as when birth control increases

welfare but reduces numbers of offspring. In the time spans of actual lives, stable goals must be shaped by patterns of reward. As the authors say, "human goals … can be any arbitrary notion of, e.g., utility or wellbeing…" (target article, Box 2).

(2) The authors cite recent brain imaging that has confirmed common-currency theories of how motives compete to determine choice (Coddington & Dudman, 2019; Levy & Glimcher, 2012), which clear the way for a reward-based economy of all mental life (Silver, Singh, Precup, & Sutton, 2021). However, the target article describes motives' operation only for the case of addiction, where a misguided reward signal leads to proxy failure (target article, sect. 4.1): This jibes with a now-standard account of drug addiction (Volkow, Wise, & Baler, 2017), but leaves unclear how the addiction case should be distinguished from routine reward-based goal setting. I propose that addictive proxies should be called hijackers when they lead to preferences that are only temporary; if they were stable they could not be distinguished from normal goals except by an eventual negative effect on evolutionary fitness (a view attributed to Becker & Murphy, 1988, target article, Box 2). A goal in the target article's terms would be an objective that survives temporary preferences.

Two mechanisms for temporary preference have been widely proposed: (a) The discounting of future outcomes in hyperbolic or similar curve (Green & Myerson, 2018) such that nearby events are overvalued; and/or (b) a temporary burst of rewarding power (Loewenstein, O'Donoghue, & Bhatia, 2015), for instance by emotional arousal or a drug. (a) Hyperbolic discounting is a robust finding both in visceral realms (as with nonhumans and young children) and deliberative activities (as in planning for retirement or dealing with global warming). However, when reported by subjects for the valuation of distant events, discounting raises the unexplored question of how quantitative expectations of such events are formed (Ainslie, 2023, pp. 19–22; Rick & Loewenstein, 2008). (b) Most rewards are enabled by appetite, the presence of which is assumed when people evaluate them at a distance. You may not be willing to pay more for a food if shopping while hungry. But some portion of future appetites seems not to be anticipated, especially when considering options that would be preferred only temporarily (Ariely & Loewenstein, 2006; Badger et al., 2007). Perhaps people avoid considering some appetites so as not to arouse them. The partial neglect of future-aroused appetite, which makes a reward curve spike upward in its presence, is also an effect that needs exploration. In any case, some mechanism of temporary preference is necessary to distinguish addictive proxies from stable goals.

(3) "Many abstract goals cannot be observed directly" (target article, sect. 3.2) – likewise distant future goals – so they invite the creation of proxies. Such creation accords with many authors' suspicion that belief is a reward-seeking activity, and even with behaviorist Howard Rachlin's radical proposal that "mental states (including sensations, perceptions, beliefs, knowledge, even pain) are… patterns of overt behavior" (2012, pp. 3–4). That is, beliefs are incentivized by their capacity to produce reward, which may happen entirely within the agent's imagination. Freed from the need for predicting external rewards, proxies may compete like fiat currencies, needing only to be protected from "inflation" by their unique "legibility" (target article, sects. 3.2 and 3.3), which thus overshadows their "ability to predict the future" as their main claim to selection.

A proxy that is a good story may turn into a goal in its own right. It may take the form of a proposition: Eating this food – or avoiding it – is morally good; there is a conspiracy to corrupt our children; Matisse produced the painting I bought. But a proxy that does not predict facts must survive through avoiding inflation by not just legibility but also *singularity* – standing out from alternative legible proxies by such factors as a unique logical argument, a parsimonious explanation, a rare coincidence, or endorsement by an authority (discussed in Ainslie, 2013, 2017, pp. 178–184, 2023, pp. 19–22). This property is important not just to psychology, but also to recent proposals that economics recognize the utility of abstract goals (Bénabou & Tirole, 2016; Loewenstein & Molnar, 2018). The self-selecting potential of proxies in neuroscience sets them apart from the other kinds of proxy the authors describe.

Financial support. This work was supported by the Department of Veterans Affairs Medical Center, Coatesville, PA, USA. The opinions expressed are not those of the Department of Veterans Affairs or of the US Government.

Competing interest. None.

References

Agarwal, D. P., & Goedde, H. W. (1989). Human aldehyde dehydrogenases: Their role in alcoholism. *Alcohol*, *6*, 517–523.

Ainslie, G. (2013). Grasping the impalpable: The role of endogenous reward in choices, including process addictions. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *56*, 446–469. doi:10.1080/0020174X.2013.806129, http://www.tandfonline.com/eprint/8fGTuFsnfFunYJKJ7aA7/full

Ainslie, G. (2017). *De gustibus disputare:* Hyperbolic delay discounting integrates five approaches to choice. *Journal of Economic Methodology*, *24*(2), 166–189. http://dx.doi.org/10.1080/1350178X.2017.1309748

Ainslie, G. (2023). Behavioral construction of the future. *Psychology of Addictive Behaviors*, *37*(1), 13–24. doi.org/10.1037/adb0000853

Ariely, D., & Loewenstein, G. (2006). The heat of the moment: The effect of sexual arousal on sexual decision making. *Journal of Behavioral Decision Making*, *19*(2), 87–98.

Badger, G. J., Bickel, W. K., Giordano, L. A., Jacobs, E. A., Loewenstein, G., & Marsch, L. (2007). Altered states: The impact of immediate craving on the valuation of current and future opioids. *Journal of Health Economics*, *26*(5), 865–876.

Becker, G., & Murphy, K. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*, 675–700.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *The Journal of Economic Perspectives*, *30*(3), 141–164.

Coddington, L. T., & Dudman, J. T. (2019). Learning from action: Reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, *104*(1), 63–77. https://doi.org/10.1016/J.NEURON.2019.08.036

Green, L., & Myerson, J. (2018). Preference reversals, delay discounting, rational choice, and the brain. In J. L. Bermúdez (Ed.), *Self-control, decision theory, and rationality: New essays* (pp. 121–146). Cambridge University Press.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/J.CONB.2012.06.001

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167.

Loewenstein, G., O'Donoghue, T., & Bhatia, S. (2015). Modeling the interplay between affect and deliberation. *Decision*, *2*(2), 55–62.

Rachlin, H. (2012). Making IBM's computer, Watson, human. *The Behavior Analyst*, *35*(1), 1–16.

Rick, S., & Loewenstein, G. (2008). Intangibility in intertemporal choice. *Philosophical Transactions of the Royal Society B*, *363*, 3813–3824.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Volkow, N. D., Wise, R. A., & Baler, R. (2017). The dopamine motive system: Implications for drug and food addiction. *Nature Reviews Neuroscience*, *18*(12), Article 12. https://doi.org/10.1038/nrn.2017.130

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

An updated perspective on teleonomy

Jonathan Bartlett

Department of Theoretical Biology, The Blyth Institute, Tulsa, OK, USA
https://www.blythinstitute.org/

Corresponding author:
Jonathan Bartlett;
Email: jonathan.l.bartlett@gmail.com

doi:10.1017/S0140525X23002868, e69

Abstract

The analysis of proxy failure given by John et al. provides a good starting point for interdisciplinary discussions. Here, the discussion of teleonomy is extended and updated to include more recent discourse on the topic.

In the analysis of proxy-based control systems (or control systems in general), one needs to determine a goal for the system. However, ascribing goals to material systems introduces a set of thorny philosophical problems about how to analyze such systems objectively.

To handle this problem, the authors of the target article use teleonomy as a type of causal explanation, which may be unfamiliar to some. In systems in which purpose is reified into a mechanism, teleonomy allows us to talk legitimately about the purposes in such systems without having to decide whether a purpose is "really" there. This allows for discussions of teleology in places where purposes seem objectively clear but it is not clear that the purposes were the result of conscious choice. This was originally developed by Pittendrigh and later by Mayr to discuss the behavior of organisms, which, in their view, exhibited clear purpose-driven behavior despite those behaviors not having been designed for a purpose (Mayr, 1961; Pittendrigh, 1958). Thus, whether from accident, intentionality, or some other cause, teleonomy provides a framework to analyze purpose in reified mechanisms. Thus, the usage of teleonomy allows John et al. to discuss the structural role of purpose across a wide variety of mechanisms that do not share similarities in their origin.

However, the discussion of teleonomy, although appreciated, falls short in two main areas. The first issue is that the authors do not address how one identifies purposes objectively in teleonomic systems. Without clear methodological principles, an investigator is in danger of assigning ad hoc purposes to systems. Thankfully, the philosophical literature provides some help. Mossio and Bich (2017) provide a solid means for practitioners to objectively identify at least some goals. They say that we can objectively classify a process as goal-directed if that goal requires the system to expend energy to accomplish and entails some aspect of maintaining the causal closure of the system.

The second area of concern is that, although the authors actively work with the concepts of teleonomy, they seem to be unaware of the progress made in the relationship between teleonomy and evolution over the past decade (Corning, 2014). Although teleonomy was originally thought to not apply to evolution itself, modern work in teleonomy has actually emphasized its role in evolution (Corning et al., 2023; Bartlett, 2023). However, the perspective that the authors of the target article take on evolution is decidedly selectionist, and, although it pays some lip service to nonselectionist thinking, it does not actually incorporate any of it into the main body of work. The extended evolutionary synthesis, which has been gaining prominence in biological thought in recent years, is largely a phenomena unified by evolutionary teleonomy (Bartlett, 2017). I think an analysis of proxy failure in evolutionary dynamics informed by modern approaches to evolution would be informative.

Additionally, the selectionist bent of the authors leads them to attribute too much causal power to natural selection whether in markets or ecology. Selection is an eliminative force, both in economics (market selection) and in biology (natural selection). Selection cannot create or fix any proxies either in the market or in biology. At most, it can eliminate bad proxies or reward good ones that already exist. Proxy creation requires positive action, and thus market selection cannot add to the increasing complexity of market forces.

Interestingly, the authors problematically equate opposing meanings of "selection" in Appendix 1. Natural selection is specifically at odds with intentional selection, yet John et al. equate "selection of action" with "selection of traits." Selection of action is an intentional process where the purposeful entity selects steps toward a goal. Natural selection is an unintentional process where bad outcomes block further elaboration of failed systems. These are fundamentally different because, in selection of action, something can be selected *prior* to the favorable outcome, in vision of it occurring, because the functional goal is "in mind." In natural selection, anything that is not presently functional is selected against, even if it is "on the way" to being functional, because natural selection cannot see future function, only present function.

This confusion creates ambiguity about what sort of causal mechanisms the authors are even referring to when discussing selection. In all, I think that the selectionist interpretation of evolutionary dynamics and its confusion with intentional selection has caused the authors to attribute to natural selection (both in the market and in nature) much more than it is capable of delivering.

This also opens up a question for future research. If selection is not creating proxies, what is? What are the teleonomic sources and/or requirements for proxy creation? How are levels of goals identified and created in such systems? Are lower-level goals inferred from higher-level ones (as is usually the case in conscious systems) or vice versa? How can such systems detect and accommodate proxy failures? Recognizing the function of teleonomy in creating these proxies will assist in asking the right questions, and perhaps also assist in building systems that are as robust as biological ones.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Bartlett, J. (2017). Evolutionary teleonomy as a unifying principle for the extended evolutionary synthesis. *BIO-Complexity*, *2017*(2), 1–7.

Bartlett, J. (2023). Causal Capabilities of Teleology and Teleonomy in Life and Evolution. *Organon F*, *30*(3), 222–254.

Corning, P. (2014). Evolution "On purpose": How behaviour has shaped the evolutionary process. *Biological Journal of the Linnean Society*, *112*(2), 242–260.

Corning, P., Kauffman, S. A., Noble, D., Shapiro, J. A., Vane-Wright, R. I., & Pross, A. (2023). *Evolution "On purpose": Teleonomy in living systems*. MIT Press.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*, 1501–1506.
Mossio, M., & Bich, L. (2017). What makes biological organisation teleological? *Synthese*, *194*, 1089–1114.
Pittendrigh, C. S. (1958). Adaptation, natural selection, and behavior. In A. Roe and G. G. Simpson (Eds.), *Behavior and evolution* (pp. 390–416). Yale University Press.

Animal welfare science, performance metrics, and proxy failure

Heather Browning[a] and Walter Veit[b]

[a]Department of Philosophy, University of Southampton, Southampton, UK and [b]Department of Philosophy, University of Reading, Reading, UK
https://www.heatherbrowning.net/
wrwveit@gmail.com, https://walterveit.com/

Corresponding author:
Heather Browning;
Email: DrHeatherBrowning@gmail.com

doi:10.1017/S0140525X2300300X, e70

Abstract

In their target article, John et al. make a convincing case that there is a unified phenomenon behind the common finding that measures become worse targets over time. Here, we will apply their framework to the domain of animal welfare science and present a pragmatic solution to reduce its impact that might also be applicable in other domains.

There is a widespread finding across different domains that the emphasis on measures can detract from the primary objectives these measures were initially designed to achieve. John et al. have examined the shared characteristics of these findings and argue that there is a unified phenomenon here that they term "proxy failure." They argue that whenever a regulator pursues a goal that involves incentivizing other agents based on their performance of a proxy measure, an inherent incentive or force emerges that causes the measure to diverge from its intended goal (p. 45). Here, we will apply their framework to the domain of animal welfare science and present a pragmatic solution to reduce its impact; insights that might also be applicable in other domains.

Animal welfare science is a relatively new discipline, having come into existence through the late 1970s and early 1980s and there was a strong pressure to be seen as a serious scientific discipline. With the behaviourist tradition still strong, talk of mental states of animals was often seen as unscientific. In order to gain more respect as a serious subject of research, animal welfare science aligned itself primarily with veterinary science, and physiology more generally. This meant that the welfare indicators chosen were those physiological indicators that were easy to measure "objectively" (Browning, 2022b; Browning & Veit, 2023). Given the difficulties in selecting and validating indicators, and the strong social and economic incentives at play within the science, ethics, and policy of animal welfare, this provides a perfect setting for the process of proxy failure described by John et al.

Indeed, there has been a focus on indicators that measured stress, such as changes in heart rate, and levels of glucocorticoids in the blood, faeces, or other tissue. Stress markers were seen as valid indicators of animal welfare through the lens of the dominant definitions of welfare at the time, which focused on how well an animal was "coping" with its situation (Broom, 1986). Stress is a measurable bodily state that can be discussed without reference to the "unscientific" mental experiences of animals and thus was seen as a suitable way of measuring welfare. We contend that early entrenchment of these stress-based welfare proxies, and ongoing social and economic incentives for their use, has led to precisely the problem John et al. describe (i.e., incentives to select for reductions in stress have led to the measures becoming worse proxies for the target state of welfare), though one that we think there are some practical means of offsetting.

Using their framework, proxy failure can be seen as an outcome of a system in which *a regulator* with a *goal* uses a *proxy* to *incentivize* or select *agents* in service of this goal. The regulator provides feedback (e.g., rewards/punishment) to the agents, adjusted based on performance as perceived through the proxy and incentivizes the agents (either actively or passively) to increase production of the proxy itself rather than the goal. In the animal welfare example we've described, the goal is broadly to improve the lives of animals. The proxies are the stress markers we have discussed. The regulatory system can be seen as the general set of social, institutional, and economic systems that impact treatment of animals by industry. Although there are specific regulators at play in different parts of the animal industry, we think that the best view is the broader systemic one, as there is no single regulator that is solely responsible for the proxy failure effect. The agents are then the animal welfare scientists, industry professionals, and policy makers, who respond to calls for improved animal welfare. The incentives are the social and economic conditions that reward or punish agents within the animal industry, including ethical concern, consumer decision making, and social licence to operate. As this feedback is provided primarily on the basis of stress-based measures, a reduction of stress (or, in the indicators used to mark it) becomes more important than a broader consideration of what is or is not good for welfare. This can be seen in the myriad studies that look at effects of changes in conditions on the glucocorticoid levels of animals, and make recommendations on this basis, without a deeper discussion of what these levels actually represent.

As the authors note, most proxies are merely approximation of their goals, and the more complex the system, the more imperfect the approximation is likely to be. Although there is no strong consensus on the best way to define welfare, there is widespread agreement that it is a highly complex state made up of multiple interacting components. The view of welfare as consisting in subjectively experienced mental states is becoming increasingly dominant (Browning, 2022a, 2023). Although stress is closely linked to welfare, and comprises part of welfare, it is not itself a complete representation of welfare. Over the past decade, there has been an increasing focus on positive welfare – the capacity of animals to experience positive experiences, rather than just the absence of negative states such as pain (Yeates & Main, 2008). An excessive focus on stress measures is partly why positive welfare was overlooked for so long.

To deal with this problem, animal welfare scientists have relied upon additional proxies that are better able to capture the aspects of welfare that stress-based proxies overlook could serve as appropriate constraints (Browning, 2020; Veit & Browning, 2020). Although the addition of new or additional proxies is identified by the authors as part of the "proxy treadmill," we think that this is not necessarily a problem, as it is unlikely to be an ongoing cycle of never-ending modification. Rather, careful selection of a few proxies that represent different aspects of welfare could lead to stabilization through their combined direct causal link to the target state. Similarly, we are confident that other fields will be able to partially overcome proxy failure by relying on a variety of different measures without becoming too attached to any single one.

Competing interest. None.

References

Broom, D. M. (1986). Indicators of poor welfare. *British Veterinary Journal*, *142*(6), 524–526.

Browning, H. (2020). If I could talk to the animals: Measuring subjective animal welfare (PhD thesis). Australian National University. https://openresearch-repository.anu.edu.au/handle/1885/206204

Browning, H. (2022a). Assessing measures of animal welfare. *Biology and Philosophy*, *37* (36), 1–24. https://doi.org/10.1007/s10539-022-09862-1

Browning, H. (2022b). The measurability of subjective animal welfare. *Journal of Consciousness Studies*, *29*(3–4), 150–179. https://doi.org/10.53765/20512201.29.3.150

Browning, H. (2023). Validating indicators of subjective animal welfare. *Philosophy of Science*, *90*(5), 1255–1264. https://doi.org/10.1017/psa.2023.10

Browning, H., & Veit, W. (2023). Studying animal feelings: Bringing together sentience research and welfare science. *Journal of Consciousness Science*, *30*(7–8), 196–222. https://doi.org/10.53765/20512201.30.7.196

Veit, W., & Browning, H. (2020). Perspectival pluralism for animal welfare. *European Journal for Philosophy of Science*, *11*(9), 1–14. doi: https://doi.org/10.1007/s13194-020-00322-9

Yeates, J. W., & Main, D. C. J. (2008). Assessment of positive welfare: A review. *The Veterinary Journal*, *175*(3), 293–300.

Proxies, heuristics, and goal alignment

Bruce D. Burns

School of Psychology, University of Sydney, Sydney, NSW, Australia
http://sydney.edu.au/science/people/bruce.burns.php

Corresponding author:
Bruce D. Burns;
Email: bruce.burns@sydney.edu.au

doi:10.1017/S0140525X2300290X, e71

Abstract

Decision-making heuristics rely on proxies so the elements of heuristics appear to map well to the elements of proxies identified by John et al. However, unlike proxy failure, heuristics do not fail because of feedback. This may be because for successful heuristics the goals of regulators and agents are aligned, but this is not the case for proxy failure.

Kahneman (2011) defined the heuristics used in decision making as substituting a difficult to answer question with an easy question to answer. For example, the availability heuristic substitutes the hard question of how large a category is or how frequent an event is, with the simpler question of how easily instances come to mind. In John et al.'s terms, the answer to such a simpler question is a proxy for the harder question. Thinking of heuristics as proxies provides a different perspective on decision-making heuristics and on proxy failure. John et al. invite the reader to identify other phenomena that have similar characteristics to proxy failure; so what can examining heuristics in terms of proxy failure tell us about both?

John et al. identify the elements of proxies and we can map these to heuristics. For the availability heuristic the regulator would appear to be a person's decision-making system, the goal is to determine frequency, the agent is a person's memory, and the proxy is ease of bringing instances to mind. The type of fast and frugal heuristics found in Gigerenzer and Todd (1999) also map to these elements. For example, they write about a successful heuristic for determining treatment of heart attack patients that asks at most three yes/no questions related to tachycardia, age, and blood pressure. If a hospital requires doctors to use this treatment heuristic, then in terms of proxies, the hospital is the regulator, the goal is saving lives, the agent is a doctor, and the proxy is the rule. (Although John et al. say that proxies are typically expressed as scalars there does not seem to be any conceptual reason why proxies could not be simple rules.)

The three limitations on regulator and agent that John et al. point to also apply to heuristics. First, there is restricted legibility of goals, because the success of heuristics is often difficult to directly observe. Second, they often make imperfect predictions because of their own limitations and because the world does not supply all the information needed. Instead they satisfice (Simon, 1955). Third, there is a necessity to choose, which is why we have heuristics.

Decision-making heuristics by definition are not guaranteed to succeed. So do they fail for the same reasons as proxies? John et al. suggest that proxy failure occurs under two conditions:

(1) There is regulatory feedback based on the proxy that has consequences for agents and
(2) the system is sufficiently complex such that there are multiple paths to the proxy that are partially independent of the goal.

The second condition appears to be necessary for both proxy and heuristic failure. For example, the body's response to heart attack can be complex so the treatment heuristic referred to above can fail because there is more than one reason why blood pressure is high. However, the first condition suggests that proxy failure is dynamic in a way that heuristic failure is not normally. For example, Goodhart's law seems to apply when the proxy starts as a good indicator of the behavior it is designed to measure, but becomes less effective as the agents learn how to game the system. Heuristics such as availability may fail, but they do not appear to degrade over time. Heuristics can fail because of flaws in the proxy they use or when the environment is manipulated to trip them up, such as when excessive media coverage of an event leads the availability heuristic to draw incorrect conclusions about the frequency of an event. Famously Tversky and Kahneman (1974) established the existence of heuristics by demonstrating conditions under which they fail, but presumably the heuristics that we persist with are ones that often succeed. If a heuristic degraded because of its use then it is likely it would fade from use in a way that common heuristics have not.

What often appears to lead to the regulatory feedback producing proxy failure is that there is a divergence of the goals of the regulator and the agent. In the example of the rat plague that opens John et al.'s paper, the rat tails may have worked reasonably well as a proxy for reducing the rat population if the residents of Hanoi had cared as much about reducing that population as the colonial officials did. Instead, the officials' goal of reducing the number of rats ran up against the residents' stronger goal of preserving a steady source of income. Successful heuristics may be successful partly because they avoid such divergence of the regulator and agent goals. For example, the memory system does not have goals that undermine the decision-making system's goals so availability can succeed; and the hospitals and doctors presumably have somewhat similar goals in wanting to treat people who will benefit the most, so the treatment heuristic succeeds.

Looked at this way, proxies are an attempt to bring the agent's goals into alignment with the regulator's goals. If the regulator's and the agent's goals are difficult to align then the proxy is in effect an attempt to impose the regulator's goals on the agent, and thus it is likely to fail if the agent has means to resist this. This suggests that an important factor determining how successful a proxy will be is how well aligned the goals of the regulator and agent are.

Examining how the phenomenon of proxy failure relates to decision-making heuristics can broaden the scope of John et al.'s analysis and says something about both heuristic and proxy failure. For heuristics, this comparison makes clear something that is not otherwise readily apparent, that the success of heuristics depends in part on the goals of the regulator and the agent being aligned, or at least not in conflict. For proxy failure this comparison suggests that failure may sometimes be less because of characteristics of proxies than to them being used to impose the regulator's goals onto the agent. So the solution to proxy failure may not be a better proxy but instead be better alignment of goals.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*, 99–118.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science (New York, N.Y.)*, *185*, 1124–1131.

On abstract goals' perverse effects on proxies: The dynamics of unattainability

José-Miguel Fernández-Dols

Facultad de Psicología, Universidad Autónoma de Madrid, Madrid, Spain

Corresponding author:
José-Miguel Fernández-Dols;
Email: jose.dols@uam.es

doi:10.1017/S0140525X23002911, e72

Abstract

Proxies should not be classified as failures or successes because, in most cases, they are impossible translations of abstract, polysemous goals to supposedly univocal concrete measures. The "success" or "failure" of a proxy does not depend on its actual accuracy as a valid indicator of goal attainment, but on a social system's willingness to maintain an illusion of its accomplishment.

One of the central arguments of John et al.'s article is that proxy failure puts in jeopardy the survival not only of all kinds of human organizations but also of organisms. Their point is intellectually seductive for its rational approach: Poorly designed proxies that are not capable of guaranteeing the accomplishment of the goals for which they were created lead social and biological systems to underperform while regulators and agents play their cards. Regulators create tighter constraints, and agents "game" the proxies created by the regulators. Eventually, an entity of inherent and superior rationality, for example the market or evolution, would get rid of these systems through a definitive process of selection.

This view of human affairs ignores the fact that rationality is not humans' strongest suit. What leads to a questionable relation between a goal and its proxy is not, in many relevant cases, the proxy, but the goal. "Proxy failure" dynamics is probably a good conceptual approach to physically highly testable goals (e.g., the three standard tests of death as a proxy for an individual's death), or goals and proxies embedded in systems ruled by constitutive norms (e.g., a checkmate counts as triumph in the world of chess, but a "maybe" does not count as getting married in a wedding ceremony). But when the goals become more abstract (and abstraction is a salient feature of institutional and organizational goals), the goals' "operationalization" is always far from ideal.

For this reason, goals such as liberty, justice, wealth, hygiene, safety, or even sanctity have been translated into a clearly diverse repertory of proxies. The problem is not these proxies, but rather the inherent impossibility of translating an idea into a univocal, exclusive actual norm or standard, and also the human tendency to assume that the designated norm or standard is the only and definitive proxy for that idea. When one reads (Vann, 2003) the details of the rat massacre in colonial Hanoi with which John et al. open their article, it becomes clear that tallying dead rats was not a faulty proxy – it was just a part of a repertory of exploitative practices aimed at constructing an ideal new France in Asia. French sewers in a tropical climate zone became a perfect vivarium for rats, and the failure of the attempt to control their population growth by sending natives to the filthy, infectious sewers to bring out dead rats was not because of a faulty proxy of the goal of killing rats but the consequence of an impossible goal. And the most plausible explanation of the natives' "gaming" of their task is not a manifestation of rational, amoral self-interest, but a way of making everyone happy amidst an impossible-to-fulfill normative framework (to install France in Vietnam). Fernández-Dols et al. (2010) showed that the apparent hypocrisy of participants in an experiment in which they could "game" the experimenter about the outcome of flipping a coin (thus earning some underserved money) was a consequence of the participants' perception of the experimenter's instructions as an arbitrary imposition; "gaming" did not happen when the instructions were clear and procedurally fair.

But even in examples such as that of colonial Hanoi, the social system does not necessarily collapse; and if it collapses is not necessarily because of the ways in which some of its core goals were translated into proxies, but typically because of the difficulties of providing an increasingly complex social system (a transplanted France) with the amount of energy needed for its survival (e.g., Tainter, 1988).

Why do systems not collapse even if the activities designed as proxies for some of its core goals are not working? The answer, for a social psychologist, is simple: Because we can create an *illusion of accomplishment* (Festinger, Riecken, & Schachter, 1956), and the social system can last for years, decades, or even centuries maintained by these illusions (Fernández-Dols, 2002). Such illusions can be protected by resorting to force or fraud, but they are much more effective when the level of abstraction of the goal itself forces the system to dictate arbitrary proxies that are *cognitively* *legitimated* as a moral or religious belief, or as a secular tradition. The more abstract these illusory core goals are, the better.

A prototypical example is religion. A recurrent theme in many contemporary movies about the Mafia is the criminals' honest respect of Catholic rites (such as the equally honest respect of Muslim rites by jihadists). This blatant failure of the faithful practice of religious rituals to attain the sanctity of their practitioners could be considered by John et al. as an ultimate, extreme example of failed proxies. We do not need to watch mafia movies to know that respecting religious rites and precepts does not guarantee sanctity. But churches last for thousands of years and do not seem to be at risk of disappearing in the foreseeable future.

Space limitations prevent me exploring examples of other social structures that last for long historical periods despite the obvious flaws in their concrete translations of abstract goals. For example, the stock market as a translation of Adam Smith's invisible hand (Shiller, 2000), or contemporary democracies as a translation of an ideal of tolerance (Levitsky & Ziblatt, 2018).

The problem is not, in many cases, the quality of the proxy as a valid construct of a goal, but the goal itself. My point is that goals are frequently resistant, for different reasons, to any accurate representation through a valid proxy. And, most importantly, untranslatable goals, and their corresponding imperfect proxies, do not compromise the survival of the social systems that legitimate these goal–proxy tandems. In some cases, these loose tandems are a guarantee of the longevity of the system itself, because imperfect proxies allow people to survive goals that otherwise would be too demanding or even inhuman, as unfortunately many social large-scale political experiments have illustrated.

Financial support. This work is part of project: Citizen perception of the elites' moral discourse and is supported by EU-UAM 2022/00085/001.

Competing interest. None.

References

Fernández-Dols, J. M. (2002). Perverse justice and perverse norms: Another turn of the screw. In M. Ross & D. T. Miller (Eds.), *The justice motive in everyday life* (pp. 79–90). Cambridge University Press.

Fernández-Dols, J. M., Aguilar, P., Campo, S., Vallacher, R., Janowsky, A., Rabbia, H., … Lerner, M. J. (2010). Hypocrites or maligned cooperative participants? Experimenter induced normative conflict in zero-sum situations. *Journal of Experimental Social Psychology*, *46*, 525–530. https://doi.org/10.1016/j.jesp.2010.02.001

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. University of Minnesota Press. https://doi.org/10.1037/10030-000

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Penguin.

Shiller, R. J. (2000). *Irrational exuberance*. Princeton University Press.

Tainter, J. A. (1988). *The collapse of complex societies*. Cambridge University Press.

Vann, M. G. (2003). Of rats, rice, and race: The great Hanoi rat massacre, an episode in French colonial history. *French Colonial History*, *4*(1), 191–203. https://doi.org/10.1353/fch.2003.0027

Subjective and objective corruption of intuition and rational choice

David Haig

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

Corresponding author:
David Haig; Email: dhaig@oeb.harvard.edu

doi:10.1017/S0140525X23002819, e73

Abstract

A societal shift has occurred toward making impactful decisions on the basis of objective metrics rather than subjective impressions. This shift is commonly justified by claims that we should not trust subjective intuitions. These are often unjust and thereby corrupt. However, the proxies used to make objective decisions are subject to a different form of corruption, characterized as proxy failure.

The themes of this commentary are tensions between subjective and objective judgment, between quality and quantity, between intuition and justified choice, between trust and accountability. Scalar proxies are increasingly employed to measure our performance. We are subject to continuous assessment and asked to justify our decisions when we evaluate others. Our choices are not taken on trust. At the same time, we rely more and more on artificial intelligences that make decisions for reasons that are opaque, both to human users and to the algorithms themselves. In order to gain our trust, these intelligences are asked to justify their intuitions (so-called "explainable AI"). A need to explain the reasons for their choices to others is the beginning of a kind of self-awareness in which artificial intelligences attempt to understand their own decisions based on proxies of their own internal processes.

Partisans of intuitive decision making reason that qualitative choices are best based on subjective feeling because quantitative measures mislead and distort. We should trust our intuition. Defenders of objective metrics feel that subjective evaluations are discriminatory, easily corrupted by criteria of choice that should be irrelevant to the decision at hand. We cannot trust first impressions.

These two modes of decision making sometimes result in different choices. Does one of them reliably make better decisions than the other? The question implies the existence of some metric on which one set of choices can be judged "better" than the other. The conceptual difference between subjective and objective choice may be less than partisans think. John et al. argue that the brain uses proxies to determine its own priorities. The opacity of an intuitive decision is a reflection of the limited insight we have into our internal use of metrics in choice.

Scalar metrics project multidimensional spaces onto lower-dimensional criteria of choice and promote uniformity of regulated behaviors. Legibility of the criteria of judgment favors working to rule, but there may be more striving for excellence when an agent does not know how their work will be judged. Monet and Picasso did not paint by numbers. Quantitative measures seem unsuited to evaluation of creative work where goals are underdetermined. A connoisseur may not know what they are looking for until they have seen it. Judgments of taste are typically intuitive rather than objectively justified.

A conceptual distinction can be made between evaluations of agents for bonuses and jobs. The problem to be solved in the design of bonuses is what incentives will *motivate* the best work *before* evaluation. Agents need to know the general goal but too much guidance from regulators about what proxies will be used in evaluating performance may undermine achievement of the goal. The problem to be solved at a job interview is what proxies will best *predict* the quality of work *after* evaluation. A less-capable candidate may succeed if they perform a more convincing impersonation of the qualities being sought.

A subject's power to make decisions that others must accept on trust is a measure of the subject's sovereign agency. Being "called to account" situates one in the relation of an untrusted subject of a

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

suspicious sovereign. (The etymological associations of *subject* and *agent*, connoting as they do both independent and dependent action, are complex. An agent works for another, and a subject is ruled by a sovereign, but I am a grammatical subject who possesses more free agency the less I need to justify my choices to others.)

In my role as an evaluator of student work, I am locked in a power struggle with students over the "subjectiveness" of my evaluations. Less-able students want to limit, I to defend, my freedom of judgment. They want me to justify their grades by showing how they conform to detailed rubrics that provide predefined metrics on which grades will be based. Administrators judge the performance of faculty by student evaluations that are contingent on the grades that we give. Faculty collude with students to obtain good evaluations for good grades and, by this process, grades are degraded.

The tension between objective metrics and subjective impressions reflects a tension between different risks of corruption. Greater legibility of the criteria of choice increases opportunities for corrupt action by agents but reduces the opportunities for corrupt choices by regulators. If our goal is to make better and fairer decisions – an abstract desideratum that is difficult to quantify – then we need to understand the interplay between these two risks of corruption. Mitigation of corrupt behavior is likely to be a constantly moving target as agents and regulators find new ways that "fair procedures" can be gamed for personal benefit.

John et al. suggest that market selection imposes a "hard constraint on proxy failure" within corporations. The exigencies of survival (natural selection) impose similar constraints on rampant proxy failure within organisms. Markets and natural selection do not play games. They are not themselves intentional agents. It is notable that the criteria for success before their tribunals have low legibility before judgment is pronounced. Revealed preference in economics and fitness in evolutionary biology are recognized after the fact and what has worked in the past is imperfectly predictive of what will work in the future. Preference and fitness are not proxies for anything else. Natural selection favors short-term genetic advantages but is not guaranteed to promote the long-term survival of an evolutionary lineage. Regulated markets possess features of intentional design by market engineers. The tragedy of the commons is not "market failure" but a predictable outcome of unregulated reliance on the "invisible hand." Self-interest is not a reliable proxy for the collective good.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

The determinants of proxy treadmilling in evolutionary models of reliable signals

Keith D. Harris

Department of Ecology, Evolution and Behavior, The Hebrew University of Jerusalem, Jerusalem, Israel

Corresponding author:
Keith D. Harris;
Email: keith.harris@mail.huji.ac.il

doi:10.1017/S0140525X23002832, e74

Abstract

Identifying the conditions of proxy treadmilling is crucial for determining whether reliable signals can persist over time. I present a framework that maps evolutionary models of reliable signals according to their assumptions regarding the effects of Goodhart's law. This framework can explain the contrasting outcomes of different modelling approaches, and identify in which models proxy treadmilling is expected to occur.

The target article raises the possibility that costly signals (i.e., signals whose level is regulated by their cost of production) are liable to a "proxy treadmill" process (McCoy & Haig, 2020). This is at odds with the conclusions of previous reliable signalling models, specifically evolutionary models that have suggested that costly signals can persist stably (Grafen, 1990; Johnstone, 1996; Zahavi, 1975). However, it is difficult to place the assumptions of these models within the scheme presented in Section 3 of the target article, partly owing to a terminological gap. To reevaluate the conclusions of evolutionary models of reliable signals in light of the target article's synthesis, I present a framework that delineates how proxy treadmilling is constrained by the assumptions of these models.

As model parameters, we include components that are commonly included in biological signalling models, and were already introduced in Fisher's runaway selection model (1930): (i) The degree to which the observer has a preference for the signal (e.g., the peahen's preference of long tails), p; (ii) the investment in the signal (e.g., elaboration of the peacock's tail), s; and (iii) the marginal cost of the signal (e.g., the survival cost associated with longer tails), m. As opposed to signal preference and investment, marginal cost often only appears implicitly (as in Fisher's model) or as a fixed property of the signal (Biernaskie, Perry, & Grafen, 2018; Harris, Daon, & Nanjundiah, 2020). By contrast, Goodhart's law subsumes the marginal cost into what determines the signal level; "proxy optimisation" in the target article refers both to an increase in signal level and a reduction in production cost. Although it has been recognised that marginal cost could change over time, resulting in the inflation and subsequent replacement of signals – as in the "inflation hypothesis" in Zahavi and Zahavi (1999) – these dynamics have not been modelled.

The interaction between the components is based on a previous model that tracks the inflation of signals (Harris et al., 2020). We assume that preference of a signal (e.g., mate selection based on signal level) can increase signal investment, s, at rate Δs, and that this investment is constrained by the marginal cost, m. If m is sufficiently low, this runaway process leads to signal inflation (Harris et al., 2020). We also allow m to decrease at rate Δm, owing to the same pressure that increases s. We assume that preference of the signal, p, will increase or decrease, depending on the benefit of observing reliable signals, at rate Δp; importantly, preference of a partially reliable signal can also be stable (Harris et al., 2020; Johnstone & Grafen, 1993). When $\Delta s < \Delta m$, we expect signal inflation to be constrained only by Δs.

We can now map outcomes of different models using two expressions, $\Delta p/\Delta s$ and $\Delta m/\Delta s$ (Fig. 1). These expressions define two axes: (i) the position on the x-axis determines the rate at which signal preference is "optimised" relative to the inflation process; and (ii) the position on the y-axis determines the degree to which marginal cost constrains signal inflation. Figure 1 also illustrates how models can be mapped in this parameter space.

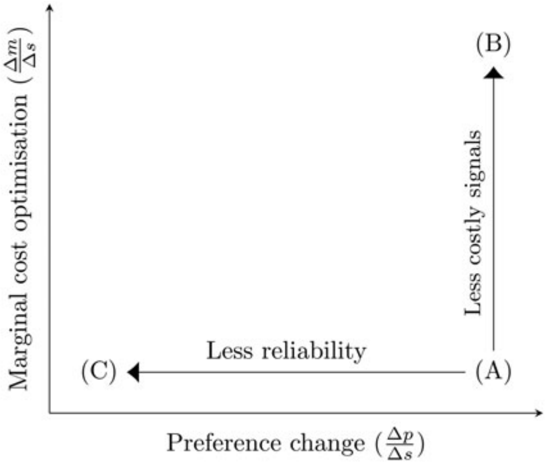
Marginal cost optimisation $\left(\frac{\Delta m}{\Delta s}\right)$
(B)
Less costly signals
Less reliability
(C)
(A)
Preference change $\left(\frac{\Delta p}{\Delta s}\right)$

Figure 1 (Harris). Mapping model outcomes based on the relationships between the rates Δp, Δs, and Δm. Arrows indicate the expected change of outcome resulting from moving along the respective axis. (A) When $\Delta p \gg \Delta s \gg \Delta m$, preference change can ignore inflated signals, while costly signals can be stable. (B) When $\Delta m > \Delta s$, signal inflation depends on Δs, and costly signals can inflate leading to a proxy treadmill, unless Δs is sufficiently low. (C) Reducing Δp relative to Δs, inflated signals can persist.

At position (A), where $\Delta p \gg \Delta s \gg \Delta m$, are models (such as the handicap principle) that assume a fixed marginal cost, and that preference change can eliminate inflated signals. However, even with these assumptions, proxy treadmilling could occur when considering perceptive error (Harris et al., 2020; Johnstone, 1994). Moving to (B) entails increasing Δm relative to Δs; this means that signal production can be optimised because of Goodhart's law in the same timescale as the change in signal investment. This also shifts the constraint on inflation to Δs. When $\Delta m > \Delta s$, signal inflation will depend only on Δs and not on m. This means that even costly signals will eventually inflate (leading to increased proxy treadmilling), whereas only signals with a sufficiently low Δs will be stable, consistent with "non-handicap" models (Számadó, 2011). Moving to (C) from (A) entails reducing Δp relative to Δs. This decreases the rate of proxy treadmilling, because inflated signals can persist for longer; indeed, it has been suggested that constraints on preference will reduce reliability (Johnstone & Grafen, 1993).

This framework demonstrates that there is a nontrivial relationship between the conclusions of evolutionary models of reliable signals and whether proxy treadmilling is expected to occur under the respective model assumptions. A fully developed model of inflation dynamics (such as Harris et al., 2020) that also allows $\Delta m > 0$ could indicate more precisely the conditions of proxy treadmilling, in terms of the axes depicted in Figure 1.

The terms introduced here can be used to differentiate between models that have reached contrasting conclusions regarding what maintains signal reliability. For instance, using this framework we find two possible explanations for "index signals": the first is that Δm is negligible, i.e., index signals have high marginal costs that are fixed, and this prevents their increase; the second is that Δs is negligible, i.e., index signals have a fixed level that cannot be manipulated. The former explanation is a costly signalling model (Zahavi & Zahavi, 1999), whereas the latter does not require costly signals to maintain signal reliability (Lachmann, Szamado, & Bergstrom, 2001). According to the above framework, testing these two explanations in specific signalling systems would involve demonstrating that the mechanism that limits Δs is unrelated to the cost of production.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Biernaskie, J. M., Perry, J. C., & Grafen, A. (2018). A general model of biological signals, from cues to handicaps. *Evolution Letters*, *2*(3), 201–209.
Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon.
Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546.
Harris, K. D., Daon, Y., & Nanjundiah, V. (2020). The role of signaling constraints in defining optimal marginal costs of reliable signals. *Behavioral Ecology*, *31*(3), 784–791.
Johnstone, R. A. (1994). Honest signalling, perceptual error and the evolution of "all-or-nothing" displays. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *256*(1346), 169–175.
Johnstone, R. A. (1996). Multiple displays in animal communication: "Backup signals" and "multiple messages". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1337), 329–338.
Johnstone, R. A., & Grafen, A. (1993). Dishonesty and the handicap principle. *Animal Behaviour*, *46*(4), 759–764.
Lachmann, M., Szamado, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13189–13194.
McCoy, D. E., & Haig, D. (2020). Embryo selection and mate choice: Can 'honest signals' be trusted? *Trends in Ecology & Evolution*, *35*(4), 308–318.
Számadó, S. (2011). The cost of honesty and the fallacy of the handicap principle. *Animal Behaviour*, *81*(1), 3–10.
Zahavi, A. (1975). Mate selection – A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214.
Zahavi, A., & Zahavi, A. (1999). *The handicap principle: A missing piece of Darwin's puzzle*. Oxford University Press.

Proxy failure and poor measurement practices in psychological science

Wendy C. Higgins[a], Alexander J. Gillett[b], David M. Kaplan[a] and Robert M. Ross[b]

[a]School of Psychological Sciences, Macquarie University, Macquarie Park, NSW, Australia and [b]Department of Philosophy, Macquarie University, Macquarie Park, NSW, Australia.
https://researchers.mq.edu.au/en/persons/wendy-higgins
alexander.gillett@mq.edu.au, https://researchers.mq.edu.au/en/persons/alexander-gillett
david.kaplan@mq.edu.au, https://researchers.mq.edu.au/en/persons/david-michael-kaplan
robert.ross@mq.edu.au, https://researchers.mq.edu.au/en/persons/robert-ross

Corresponding author:
Wendy C. Higgins;
Email: wendy.higgins@mq.edu.au

doi:10.1017/S0140525X2300287X, e75

Abstract

We argue that proxy failure contributes to poor measurement practices in psychological science and that a tradeoff exists between the legibility and fidelity of proxies whereby increasing legibility can result in decreased fidelity.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

John et al. offer insights about proxy failure across a range of disciplines (see their Table 1). We argue that this proxy failure lens can also be fruitfully applied in psychological science, where construct validity serves as a proxy for the goal of measuring unobservable psychological phenomena. Validated measurements (i.e., scores on self-report questionnaires and tests of ability) then serve as proxies for higher-order goals, such as improving clinical outcomes.

The American Psychological Association guidelines state: "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (American Educational Research Association et al., 2014, p. 21). Further, because the validity of test scores can vary depending on the properties of the sample being examined, "[b]est practice is to estimate both reliability and validity, when possible, within the researcher's sample or samples" (Appelbaum et al., 2018, p. 9). Unfortunately, there is a growing body of evidence demonstrating that studies across psychological science routinely accept measurements as valid without sufficient validity evidence (Alexandrova & Haybron, 2016; Flake, Pek, & Hehman, 2017; Higgins, Ross, Polito, & Kaplan, 2023b; Hussey & Hughes, 2020; Shaw, Cloos, Luong, Elbaz, & Flake, 2020; Slaney, 2017), including measurements used for important clinical applications such as diagnosing and treating depression (Fried, Flake, & Robinaugh, 2022).

Building on the work of John et al., we propose that the proxy failure framework highlights a key cause of inadequate construct validity evidence in psychological science: When test scores become targets, focus is diverted away from the *relationship* between test scores and the underlying psychological constructs they are intended to measure. This, we suggest, can result in divergence between test scores and psychological phenomena. For instance, tests are sometimes used in different populations without considering whether the relationship between test scores and psychological constructs holds across populations.

Consider the Reading the Mind in the Eyes Test (Eyes Test; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), which is widely used as a measure of social cognitive ability in samples drawn from many clinical and nonclinical populations and countries (Pavlova & Sokolov, 2022). Despite the near universal practice of calculating a single sum score for the 36-item Eyes Test, there are two key pieces of evidence that the structural properties of Eyes Test scores vary across samples and that the interpretation of sum scores is not always supported. First, factor analysis studies spanning multiple language versions of the Eyes Test have reported poor unidimensional model fit (e.g., Dordevic et al., 2017; Higgins, Ross, Langdon, & Polito, 2023a; Olderbak et al., 2015; Redondo & Herrero-Fernández, 2018; Topić & Kovačević, 2019), and it has even been found that the factor structure of Eyes Test scores for different ethnic and linguistic groups within the same country can vary (Van Staden & Callaghan, 2021). Second, a recent meta-analysis identified substantial variation in the internal consistency estimates of Eyes Test scores across samples, with half falling below the level conventionally taken to be acceptable (Kittel, Olderbak, & Wilhelm, 2022). Yet, Eyes Test sum scores are frequently compared between populations, with inferences drawn about relative levels of social cognitive ability.

An outstanding question when the proxy failure framework is applied to psychological science is why studies that fail to meet existing construct validity evidence reporting standards are routinely published (Flake & Fried, 2020; Slaney, 2017). As John et al. note, a proxy must be simple enough for agents and regulators to identify and understand (i.e., must be legible), so that it can "feasibly be observed, rewarded, and pursued" (target article, sect. 3.2, para. 2). However, legibility can come at a cost to fidelity: "There is a natural human tendency to try to simplify problems by focusing on the most easily measurable elements. But what is most easily measured is rarely what is most important" (Muller, 2018, p. 23). Although psychological research standards state that the construct validity proxy should be based on a variety of sources of validity evidence (American Educational Research Association et al., 2014; Clark & Watson, 2019), some sources of evidence are more legible than others. We contend that the failure to enforce best practices in construct validation can be explained in part because of the prioritisation of legibility over fidelity. This results in less legible sources of validity evidence being overlooked in a phenomenon we refer to as "proxy pruning." Unfortunately, proxy pruning can result in test scores being accepted as valid that might be deemed invalid if other sources of validity evidence were examined.

A key example of proxy pruning in psychological science is ignoring the importance of providing construct validity evidence that is derived from a psychological theory (Alexandrova & Haybron, 2016; Bringmann, Elmer, & Eronen, 2022; Eronen & Bringmann, 2021; Feest, 2020). In particular, researchers often over-rely on the psychometric properties of test scores when establishing construct validity and avoid challenging theoretical questions about how to define psychological constructs (Alexandrova & Haybron, 2016; Clark & Watson, 2019). In addition to being important in their own right, these theoretical questions can be critical to interpreting psychometric properties. Consider the use of convergent validity evidence in the well-being literature where "it can seem as if nearly everything correlates substantially with nearly everything else" (Alexandrova & Haybron, 2016, p. 1104). Some researchers have deemed that "better" measures of well-being are those that correlate more strongly with measures of life circumstances, including income, governance, and freedom. However, without having done the hard conceptual work of determining what precisely is meant by "well-being" (e.g., "happiness," "life satisfaction," "flourishing," "preference satisfaction," "quality of life"; Alexandrova & Singh, 2022) it remains unclear why correlating more strongly with these particular variables is indicative of a more valid measure of well-being.

In sum, John et al.'s proxy failure framework offers insights into poor measurement practices in psychological science. However, we argue that the problem with proxies is not only that they become targets, but that they are pruned down to be legible targets, which decreases their fidelity, leaving them more susceptible to proxy failure.

Financial support. This work was supported by an Australian Government Research Training Program (RTP) Scholarship (W. C. H.), a Macquarie University Research Excellence Scholarship (W. C. H.), and the John Templeton Foundation (R. M. R., grant number 62631; A. J. G., grant number 61924).

Competing interest. None.

References

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science*, *83*(5), 1098–1109. https://doi.org/10.1086/687941

Alexandrova, A., & Singh, R. (2022). When well-being becomes a number. In C. Newfield, A. Alexandrova, & S. John (Eds.), *Limits of the numerical: The abuses and uses of quantification* (pp. 181–199). University of Chicago Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1017/S0021963001006643

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science: A Journal of the American Psychological Society*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Dordevic, J., Zivanovic, M., Pavlovic, A., Mihajlovic, G., Karlicic, I. S., & Pavlovic, D. (2017). Psychometric evaluation and validation of the Serbian version of "Reading the Mind in the Eyes" test. *Psihologija*, *50*(4), 483–502. https://doi.org/10.2298/PSI170504010D

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Feest, U. (2020). Construct validity in psychological tests – The case of implicit social cognition. *European Journal for Philosophy of Science*, *10*(1), 4. https://doi.org/10.1007/s13194-019-0270-8

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2023a). The "Reading the Mind in the Eyes" test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment*, *30*(6), 1777–1789. https://doi.org/10.1177/10731911221124342

Higgins, W. C., Ross, R. M., Polito, V., & Kaplan, D. M. (2023b). Three threats to the validity of the Reading the Mind in the Eyes Test: A commentary on Pavlova and Sokolov (2022). *Neuroscience and Biobehavioral Reviews*, *147*, 105088. https://doi.org/10.1016/j.neubiorev.2023.105088

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the mind's eye: A meta-analytic investigation of the nomological network and internal consistency of the "Reading the Mind in the Eyes" test. *Assessment*, *29*(5), 872–895. https://doi.org/10.1177/1073191121996469

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503. https://doi.org/10.3389/fpsyg.2015.01503

Pavlova, M. A., & Sokolov, A. A. (2022). Reading language of the eyes. *Neuroscience and Biobehavioral Reviews*, *140*, 104755–104755. https://doi.org/10.1016/j.neubiorev.2022.104755

Redondo, I., & Herrero-Fernández, D. (2018). Validation of the Reading the Mind in the Eyes Test in a healthy Spanish sample and women with anorexia nervosa. *Cognitive Neuropsychiatry*, *23*(4), 201–217. https://doi.org/10.1080/13546805.2018.1461618

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from many labs 2. *Canadian Psychology=Psychologie Canadienne*, *61*(4), 289–298. https://doi.org/10.1037/cap0000220

Slaney, K. (2017). *Validating psychological constructs historical, philosophical, and practical dimensions*. Palgrave Macmillan.

Topić, M. K., & Kovačević, M. P. (2019). Croatian adaptation of the revised Reading the Mind in the Eyes Test (RMET). *Psihologijske Teme*, *28*(2), 377–395. https://doi.org/10.31820/pt.28.2.8

Van Staden, J. G., & Callaghan, C. W. (2021). An evaluation of the reading the mind in the eyes test's psychometric properties and scores in South Africa-cultural implications. *Psychological Research*, *86*(7), 2289–2300. https://doi.org/10.1007/s00426-021-01539-w

Proxy failures in practice: Examples from the sociology of science

Jakob Kapeller[a,b], Stephan Pühringer[b], Johanna Rath[b] and Matthias Aistleitner[b]

[a]Institute for Socio-Economics, University of Duisburg-Essen, Duisburg, Germany and [b]Institute for Comprehensive Analysis of the Economy (ICAE), Johannes Kepler University, Linz, Austria.
stephan.puehringer@jku.at
johanna.rath@jku.at
matthias.aistleitner@jku.at
https://www.uni-due.de/soziooekonomie/institut_en
https://www.jku.at/en/institute-for-comprehensive-analysis-of-the-economy/

Corresponding author:
Jakob Kapeller;
Email: jakob.kapeller@uni-due.de

doi:10.1017/S0140525X23002856, e76

Abstract

Following John et al., we provide examples of failing proxies that might help to contextualize the role of proxy failures in applied research. We focus on examples from the sociology of science and illustrate how the notion of proxy failure can sharpen applied analysis, if used in a way that does not obscure other dysfunctional effects of proxies.

We applaud the efforts of John et al. to develop a general theory on the dysfunctionality of proxies, by providing a clear definition for a *proxy failure* based on a "unifying mechanism," that can be observed in diverse social and biological systems. This mechanism states that optimization in complex systems creates an endogenous tendency for the proxy to diverge from the true underlying goal.

However, *proxy failures* in the spirit of John et al. are not the only way in which *proxies can fail*, mostly because there is typically only a partial overlap between proxies and intrinsic goals. This gives rise to dysfunctional practices (for a transdisciplinary review, see Braganza, 2022) even without assuming some reinforcement effect.

In what follows, we illustrate this argument with three examples from the sociology of science – research evaluation, the political economy of publishing, and the third mission in academia. In all three cases, proxies fail in some way, but only the first case represents a clear-cut example of *proxy failure* in the sense of John et al. We take two main insights from these three examples: First, *proxy failure* as developed by the authors can improve and refine applied research. Nonetheless, many relevant instances of *failing proxies* fall outside the narrower definition of a *proxy failure*. This is because not every relevant dysfunctionality of a proxy is accompanied by a feedback loop that amplifies the distance between proxy and goal.

These insights also have repercussions for how broadly applicable *proxy failure* is and what role it plays relative to other causes of *failing proxies* – this facet could be more deeply explored in further research by scrutinizing examples for *proxy failures* as supplied in Table 1 in John et al. along the lines suggested here.

Our first example relates to *proxy failure* in the context of research evaluation driven by (the number of) citations. "Cheats in citation game" (Biagioli, 2016) is an extremely well-documented phenomenon in academia and John et al. also refer to it as an example for a *proxy failure*. In this context, the universality of *proxy failures* can even be hypothesized to operate on different levels. A general example is given by the size bias that may arise in scientific evolution (Sterman & Wittenberg, 1999): large and established research fields are more attractive than small (and potentially disruptive) ones, which further inflates the relative size of the former (see Aistleitner, Kapeller, & Steinerberger, 2018). The underlying logic of preferential attachment then impacts and biases the distribution of citations, which in turn is used to evaluate researchers, departments, and journals, which, again, impacts visibility, thereby creating another feedback loop.

Our second example pertains to the political economy of scientific publishing, where profit, as a classic proxy measure of firm performance, seems particularly inadequate. Although public debate often conveys the impression that firms with high profits also show high performance, a more critical stance would also look for the actual sources of these profits. In this spirit, closer inspection suggests that profits as a proxy measure appear to be fundamentally ill-suited to the context of scientific publishers. These firms operate in an environment characterized by substantial indirect subsidies as well as monopoly rents derived from intellectual property rights and intrinsic motivations of researchers, who provide research manuscripts and peer-review services without financial reward. This combination results in disproportionately high profit margins ranging from 20 to 40%, which significantly surpasses profit margins achieved in other industries. In this context, the mismatch between proxy and underlying goals gives rise to "*corrupted practices*" (Braganza, 2022), that adversely affect the societal goal by appropriating a public good for private gain (Pühringer, Rath, & Griesebner, 2021). This prime dysfunctionality is ex-ante unrelated to a *proxy failure*. However, such a failure can be reconstructed with reference to the trend toward concentration witnessed by the scientific publishing industry in recent years. Such increasing concentration could indeed map well on John et al.'s assertion of *proxy* failure by further pushing up profit rates. However, such a pattern is arguably more difficult to identify and not necessary for recognizing that profits may be an inherently misleading proxy for firm performance.

Our third and final example relates to public impact of science. The "third mission" in academia has taken an important role in research evaluation, which typically relies on proxies, such as the number of public appearances or the citations in policy documents. Here our main concern is that these proxies hardly assess whether the consequences of some public impact are conducive to the goal of the "third mission" (defined as tackling societal challenges). Eugenicists had a huge audience in the 1920s and the jury on those famous economists and political scientists helping to implement shock therapy after the fall of the Soviet Union is supposedly still out (Pistor, 2022). Although these examples suggest that a qualitative critique of proxies is inherently necessary and that proxy competition may be instrumentalized for political ideologies (Braganza, 2022), they do not directly speak to the narrower notion of a proxy failure. Nonetheless, similar to our second example, an argument along the lines of a *proxy failure* could be made. This would require the proxy to somehow deteriorate the quality of inputs from science to society, maybe because the incentive to receive public attention may cause scientists to be less careful or sensible in their public statements.

In concluding, we note that (potentially) failing proxies are anywhere, and the instance of this commentary in Behavioral and Brain Sciences (BBS) itself provides a compelling example. Evaluative metrics such as the Journal Impact Factor (JIF) typically count citations per article. Hence, if Web of Science were to classify comments like this one as "full articles," publishing them would automatically depress the JIF of BBS. Hence, the (non)existence of open peer commentary in BBS may ultimately rather rely on some detail in the inner workings of the evaluation industry, than on its – undoubtedly intrinsic – merit for scientific advancement.

Financial support. Pühringer and Rath acknowledge support from the Austrian Science Fund (FWF, Grant Number ZK60-G27).

Competing interest. None.

References

Aistleitner, M., Kapeller, J., & Steinerberger, S. (2018). The power of scientometrics and the development of economics. *Journal of Economic Issues*, *52*, 816–834. doi:10.1080/00213624.2018.1498721

Biagioli, M. (2016). Watch out for cheats in citation game. *Nature News*, *535*, 201. doi:10.1038/535201a

Braganza, O. (2022). Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, *9*, 211030. doi:10.1098/RSOS.211030

Pistor, K. (2022, February 28). From shock therapy to Putin's war. Retrieved from https://www.project-syndicate.org/commentary/1990s-shock-therapy-set-stage-for-russian-authoritarianism-by-katharina-pistor-2022-02

Pühringer, S., Rath, J., & Griesebner, T. (2021). The political economy of academic publishing: On the commodification of a public good. *PLoS ONE*, *16*(6), e0253226. doi:10.1371/journal.pone.0253226

Sterman, J. D., & Wittenberg, J. (1999). Path dependence, competition, and succession in the dynamics of scientific revolution. *Organization Science*, *10*, 322–341. doi:10.1287/orsc.10.3.322

Dynamic diversity is the answer to proxy failure

Zeb Kurth-Nelson[a,b,*], Steve Sullivan[c,*], Joel Z. Leibo[a] and Marc Guitart-Masip[b,d,e,f]

[a]Google DeepMind, London, UK; [b]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK; [c]Department of Anesthesiology and Perioperative Medicine, Oregon Health and Science University, Portland, OR, USA; [d]Aging Research Center, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden; [e]Center for Psychiatry Research, Region Stockholm, Stockholm, Sweden. Center for Cognitive and [f]Computational Neuropsychiatry (CCNP), Karolinska Institutet, Stockholm, Sweden
sulliste@ohsu.edu
jzl@google.com
marc.guitart78@gmail.com

Corresponding author:
Zeb Kurth-Nelson;
Email: zebkurthnelson@gmail.com

doi:10.1017/S0140525X23002923, e77

*Equal contribution

Abstract

We argue that a diverse and dynamic pool of agents mitigates proxy failure. Proxy modularity plays a key role in the ongoing production of diversity. We review examples from a range of scales.

The ingredients for proxy failure are a target and an agent that optimizes for an approximation (proxy) of the target. Because the proxy is not the actual target, the behavior of the agent can become misaligned with the target (John et al.). In fact, Sohl-Dickstein (2022) points out that if proxy optimization is too efficient, it reliably becomes not only ineffective but also actively harmful. Here, we argue, from molecules to societies, that the harm of proxy failure is minimized by a diverse and dynamic population of proxies; and that periodic separation between agents forces them to both individualize and work together, leading to new solutions.

John et al. give the example of decision-making algorithms in the brain as proxies for evolutionary fitness. These proxies fail with, for example, abused drugs or excessive consumption of food. In our view, diversity in decision-making systems is a central defense against this kind of proxy failure. The hypothalamus contains a set of segregated circuits, each implementing a distinct "hard-wired" behavioral policy aimed toward one homeostatic or reproductive goal, such as feeding, drinking, or mating (Saper & Lowell, 2014; Schulkin & Sterling, 2019; Sewards & Sewards, 2003). In service of basic drives, corticostriatal circuitry also learns a more general and flexible set of goals (Balleine, Delgado, & Hikosaka, 2007; Cardinal, Parkinson, Hall, & Everitt, 2002; Frank & Claus, 2006; Saunders & Robinson, 2012). Of course, the behaviors prescribed by different goals often conflict, and the striatum can be viewed as a "parliament" dynamically arbitrating between goals (Cui et al., 2013; Da Silva, Tecuapetla, Paixão, & Costa, 2018; Graybiel & Grafton, 2015; Klaus et al., 2017; Mohebi et al., 2019). Humans in particular adopt a dizzying diversity of goals (O'Reilly, Hazy, Mollick, Mackie, & Herd, 2014; Schank & Abelson, 1977) and also synthesize new goals when existing ones are frustrated. Each goal represents a different proxy for evolutionary fitness, and they better approximate fitness when they are in balance than when an individual goal is excessively optimized. Pathological states occur when the system gets stuck on a single goal, such as in addiction or rumination.

Diversity of beliefs protects against proxy failure in the same way as diversity of goals. Every human holds many distinct beliefs. The beliefs are "separate," in that they are not required to be consistent with one another (Wood, Douglas, & Sutton, 2012), and when one is active, others are largely inaccessible (Hills, Todd, Lazer, Redish, & Couzin, 2015). Each belief (or perspective, or metaphor) is only a partial description of the world – a proxy for a broader truth. This proxy diversity serves us well. An individual with multiple perspectives on a problem is less likely to get stuck in a particular approach (De Bono, 1970; Duncker, 1945; Ohlsson, 1992), and a deep understanding of a topic means having many different perspectives available (Feyerabend, 1975; Lakoff & Johnson, 1980; Saffo, 2008; Wittgenstein, 1953). Conversely, if we attach to and optimize for a single perspective, our thinking is rigid and shallow: Optimizing too strongly for that single proxy leads to divergence from the broader truth. In the brain, a network centered on hippocampus appears to support diversity and dynamism. This network separates knowledge modularly into distinct entities and narratives (McClelland, McNaughton, & O'Reilly, 1995; Yassa & Stark, 2011). Vitally, after they are separated, the entities are then also flexibly composed together in many different ways, synthesizing new knowledge and perspectives (Buckner, 2010; Kazanina & Poeppel, 2023; Kurth-Nelson et al., 2023; O'Reilly, Ranganath, & Russin, 2022).

Just as the brain holds diverse motivations and beliefs in balance, multiagent systems such as human societies contain diverse and competing forces, which can be seen as proxies for collective welfare. There is a rich tradition of studying the conditions under which this diversity of objectives is conducive to broader success (Ostrom, Gardner, & Walker, 1994). Empirically, excess communication reduces diversity and worsens performance in human groups (Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Page, 2017). However, if individuals are allowed to spend time first working on a problem in isolation and then combine solutions, the group performs better (Bernstein, Shore, & Lazer, 2018). This example follows the general pattern that entities must first separate to diversify and gain individual stability. Then, interaction creates higher-order structures, leading to hierarchies and open-ended evolution.

Diversity plays a similar role in groups of artificial agents. Imagine an evolving population of game-playing agents, where the fitness of each individual is determined by playing paper-rock-scissors against each other. If the population loses diversity and collapses on a single strategy, such as "always play rock," then a mutation that produces the strategy "always play paper" will dominate the population. These waves of dominant strategies can go in circles through the optimization landscape, never improving overall. However, if the population is diverse, agents are forced to discover truly new solutions, an effect also documented in much more complex games (Crepinšek, Liu, & Mernik, 2013; Czarnecki et al., 2020; Leibo, Hughes, Lanctot, & Graepel, 2019; Vinyals et al., 2019).

As a final example, sexual reproduction is remarkably common (Judson & Normark, 1996; Speijer, Lukeš, & Eliáš, 2015), despite the cost of producing males and the challenge of finding mates in the vast world (Lehtonen, Jennions, & Kokko, 2012; Maynard Smith, 1978). What advantages does sex offer? A traditional view is that recombination generates diversity by exploring new combinations of genes. A fascinating extension of this theory is that recombination also forces the genes to be modular, democratizing the genome (Agren, Haig, & McCoy, 2022; Livnat, Papadimitriou, Dushoff, & Feldman, 2008; Melo, Porto, Cheverud, & Marroig, 2016; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014; Veller, 2022). A gene can't depend on the presence of another particular gene because it might disappear in the next shuffling. Instead, each gene is incentivized to function productively with any new genome it finds itself in – yielding a genetic foundation ripe for synthesis of new solutions. Although each gene is selfish and is only an imperfect proxy for the welfare of the organism, a diverse and dynamic set of genes protects against proxy failure.

In conclusion, connectedness must be balanced with periods of separation to maintain diversity and protect against proxy failures. We should be cautious about moving toward continual interconnectedness and premature exchange of information. Similarly, with rapid advances in artificial intelligence (AI), we should be cautious about concentrating intelligence in one place. Diverse AI systems should exist with different objectives and modes of operation. Troublingly, proxy failure may explain

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

the Fermi paradox – the puzzle that we don't see other intelligent life in the universe. Through Earth's history, evolutionary experiments have had opportunities to develop separately. Archaea and prokaryotic mitochondrial ancestors specialized separately for hundreds of millions of years before achieving the distinct forms that enabled fruitful endosymbiosis, fueling the explosion of multicellular complexity (Lane & Martin, 2010; Margulis, 1970; Roger, Muñoz-Gómez, & Kamikawa, 2017). However, the trend with increased intelligence is toward immediate exchange of information between entities across the planet, reducing proxy diversity, with risk of catastrophic failure (Diamond, 2005).

Acknowledgments. We thank Zora Wessely for her comments on an earlier version of the manuscript.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. Z. K.-N. and J. Z. L. are employed by Google DeepMind. S. S. and M. G.-M. have no competing interest to declare.

References

Agren, J. A., Haig, D., & McCoy, D. E. (2022). Meiosis solved the problem of gerrymandering. *Journal of Genetics*, *101*(2), 38.

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, *27*(31), 8161–8165.

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8734–8739.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, *61*, 27–48.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, *26*(3), 321–352.

Crepinšek, M., Liu, S.-H., & Mernik, M. (2013). Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Surveys (CSUR)*, *45*(3), 1–33.

Cui, G., Jun, S. B., Jin, X., Pham, M. D., Vogel, S. S., Lovinger, D. M., & Costa, R. M. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, *494*(7436), 238–242.

Czarnecki, W. M., Gidel, G., Tracey, B., Tuyls, K., Omidshafiei, S., Balduzzi, D., & Jaderberg, M. (2020). Real world games look like spinning tops. *Advances in Neural Information Processing Systems*, *33*, 17443–17454.

Da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*(7691), 244–248.

De Bono, E. (1970). *Lateral thinking* (Vol. 70). Harper & Row.

Diamond, J. (2005). *Collapse: How societies choose to fail or succeed*. Penguin Books. ISBN 9780241958681.

Duncker, K. (1945). On problem-solving. *Psychological monographs*, *58*(5), 1–113.

Feyerabend, P. K. (1975). *Against method*. Verso.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300.

Graybiel, A. M., & Grafton, S. T. (2015). The striatum: Where skills and habits meet. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021691.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54.

Judson, O. P., & Normark, B. B. (1996). Ancient asexual scandals. *Trends in Ecology & Evolution*, *11*(2), 41–46.

Kazanina, N., & Poeppel, D. (2023). The neural ingredients for a language of thought are available. *Trends in Cognitive Sciences*, *27*(11), 996–1007.

Klaus, A., Martins, G. J., Paixao, V. B., Zhou, P., Paninski, L., & Costa, R. M. (2017). The spatiotemporal organization of the striatum encodes action space. *Neuron*, *95*(5), 1171–1180.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., … Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lane, N., & Martin, W. (2010). The energetics of genome complexity. *Nature*, *467*(7318), 929–934.

Lehtonen, J., Jennions, M. D., & Kokko, H. (2012). The many costs of sex. *Trends in Ecology & Evolution*, *27*(3), 172–178.

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv*:1903.00742.

Livnat, A., Papadimitriou, C., Dushoff, J., & Feldman, M. W. (2008). A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(50), 19803–19808.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9020–9025.

Margulis, L. (1970). *Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian Earth*. Yale University Press.

Maynard Smith, J. (1978). *The evolution of sex* (Vol. 4). Cambridge University Press.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

Melo, D., Porto, A., Cheverud, J. M., & Marroig, G. (2016). Modularity: Genes, development, and evolution. *Annual Review of Ecology, Evolution, and Systematics*, *47*, 463–486.

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J.-M. T., Vinson, L. T., Patriarchi, T., … Berke, J. D. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature*, *570*(7759), 65–70.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, *1*, 1–44.

O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P., & Herd, S. (2014). Goal-driven cognition in the brain: A computational framework. *arXiv preprint arXiv*:1404.7591.

O'Reilly, R. C., Ranganath, C., & Russin, J. L. (2022). The structure of systematicity in the brain. *Current Directions in Psychological Science*, *31*(2), 124–130.

Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games, and common-pool resources*. University of Michigan Press.

Page, S. E. (2017). *The diversity bonus*. Princeton University Press.

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The origin and diversification of mitochondria. *Current Biology*, *27*(21), R1177–R1192.

Saffo, P. (2008). Strong opinions weakly held. https://saffo.com/02008/07/26/strong-opinions-weakly-held/

Saper, C. B., & Lowell, B. B. (2014). The hypothalamus. *Current Biology*, *24*(23), R1111–R1116.

Saunders, B. T., & Robinson, T. E. (2012). The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. *European Journal of Neuroscience*, *36*(4), 2521–2532.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schulkin, J., & Sterling, P. (2019). Allostasis: A brain-centered, predictive mode of physiological regulation. *Trends in Neurosciences*, *42*(10), 740–752.

Sewards, T. V., & Sewards, M. A. (2003). Representations of motivational drives in mesial cortex, medial thalamus, hypothalamus and midbrain. *Brain Research Bulletin*, *61*(1), 25–49.

Sohl-Dickstein, J. (2022). Too much efficiency makes everything worse: Overfitting and the strong version of Goodhart's law, November 2022. https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

Speijer, D., Lukeš, J., & Eliáš, M. (2015). Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(29), 8827–8834.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Veller, C. (2022). Mendel's first law: Partisan interests and the parliament of genes. *Heredity*, *129*(1), 48–55.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019). Grandmaster level in Starcraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354.

Wittgenstein, L. (1953). *Philosophical investigations*. Basil Blackwell.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, *3*(6), 767–773.

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Regulator and agent sophistication as an explanation-generating engine for proxy failure dynamics

Mihnea Moldoveanu

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Corresponding author:
Mihnea Moldoveanu; Email: Mihnea.moldoveanu@rotman.utoronto.ca

doi:10.1017/S0140525X23002881, e78

Abstract

I consider the dynamics of regulator–agent interactions in situations in which there are significant mismatches in their abilities to discern and register information and calculate and act upon successful inferences.

The proxy failure scenario of the target article posits a – social, ecological, or physiological – system monitored/controlled by a regulator that selects an observable measure proxying for the system's current state, or for the distance between the system's current state and a goal state, as a basis for allocating rewards to an agent who acts upon the system on the basis of the incentives provided. When agent and regulator have differing objectives and are capable of mutual representation and manipulation – as they often are in a human organization – their relative levels of sophistication significantly impact the dynamics of proxy failures.

Let us consider human agents in organizations such as a business or an institution. They vary in their abilities to classify, observe, predict, and manipulate their environments, which *include the regulators to whom they "report."* A salesperson can manipulate a commission-based compensation plan in which her quarterly earnings grow nonlinearly as a function of revenue generated by closing all of her sales for 1 year in a single quarter, and thus realizing a much higher "cut" than what she would have received had the closings been evenly spaced. If her quarterly commission is 10% for the first $1,000,000 sold, 15% for the next $1,000,000, and 20% for anything over $2,000,000 and she is looking at closing some $4,000,000 in sales in 1 year, then the difference between piling a year's revenue onto one quarter (commission = $650,000) and smoothing it out over each of the four quarters (commission = $400,000) is $250,000 – more than 5% of the total revenue she has brought in. The cost to the system (the business) is the additional value of the salesperson's earnings, *plus* the resulting "spikiness" of the revenue, which increases the volatility of its equity and lowers its value.

This is a *simple* example of an agent being just a little more sophisticated than her regulator. Because it is simple, it tempts us to think the regulator can repair the incentive scheme: Impose a minimum quarterly revenue, make the commission a fixed percentage of the revenue generated, use options on equity to incentivize salesperson to take actions that benefit the business as a whole, and so on. That is what we see in both studies of relational contracting and the "common law" generated by employment and compensation agreements. But, can an agent more sophisticated than the regulator "almost always" game a proxy measure for her own advantage? In our example, she could "play around" with the standard revenue recognition metrics to have more revenue counted toward her incentive plan than the company registers; or "slag" other salespeople on the same organization to their clients in order to have them become her clients; and so on.

The agent's services are retained and contracted for in the first place *because* she has a higher level of *skill in performing a task* than the regulator does. A skill is a mapping of effort and ability onto outcomes in the context of a task the skill enables the agent to carry out. Because of the structure of the regulator–agent problem, the agent has a *sophistication advantage* over the regulator: She will almost-always be able to produce outcomes that satisfy the assessments of the regulator but "game" the incentive scheme assessments drive. To make sophistication mismatch sharper, distinguish between:

- *Informational sophistication: Perception, registration, encoding, remembering.* The agent "sees more," makes more distinctions, sees more clearly, and remembers more reliably. If the agent is an expert who makes predictions about system-relevant events, a regulator may incentivize her with a "$A + B\ \log(p(e_i))$" contract (Moldoveanu, 2011), which pays her a fixed wage A and subtracts a fine proportional to the probability $p(e_i)$ if/when event e_i occurs. This scheme is "optimal" (Bernardo, 1979) provided the regulator and the agent have the same state space of events $\{e_k\}$. But defining the "right event space" is *precisely* what the regulator needs the agent for. A wily agent can articulate "event descriptions" malleable enough to encompass most of what the agent thinks the regulator can "see" (say, broad categories in a market or technology forecast), and ascribe *high* probabilities to them, which minimize her $\log(p(e_i))$ fines in spite of not making a useful prediction.
- *Computational sophistication: Calculation and behavior.* The agent "thinks more quickly," "sees further into the future," and "takes more actions." If regulators attempt to remedy proxy failures via "patches" that locally address failures of the measurement to reflect an objective for the system then the agent–regulator "game" becomes a sequence of moves in which "the quick" kill off "the slow" (Moldoveanu, 2009). In a typical CEO–Chairperson (CE–CH) game in which CH tries to oust CE for poor performance or fraudulent behavior, CE, knowing CH needs to individually persuade board members of the necessity of the move, can map out "influence networks" by which she can persuade crucial board members of the faulty reasoning or bias of CH and, using her speed advantage, marshal the board clique necessary to pre-emptively oust CH. Both inferential speed (thinking one move ahead of CH) and behavioral agility (making the requisite number of calls) count: But those are also core skills for which CE was hired.

What adjustments and recourses can a regulator access? A "sophisticated" regulator is not *as sophisticated as an agent* in the domains where he needs the agents' expertise. Rather, *she understands the dynamics of proxy failures given mismatches in sophistication*, and knows *he cannot by himself repair them.* So, he can:

- Hire, as subregulators, former agents that understand the games most agents play in a niche – as we see in the case of venture capital and private equity firms.
- Create high-powered incentives for agents (e.g., high CEO equity stakes) that allow the latter to themselves become

principals and regulators if they deliver on the results that condition their discharge.
- Hire experts (e.g., consulting firms) that help keep sophisticated agents honest in the face of temptations to realize gains at the expense of the system (shareholders).

None is fail-safe against a wily, sophisticated agent. The "system" (of "capital preservation") ultimately relies for resilience on the desire of agents to "one day" themselves become principals and regulators (as owners of capital) – and not just to "win the game."

Financial support. This work was supported by the Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto.

Competing interest. None.

References

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, *7*, 171–187.

Moldoveanu, M. C. (2009). Thinking strategically about thinking strategically: The computational structure and dynamics of managerial reasoning. *Strategic Management Journal*, *30*, 737–763.

Moldoveanu, M. C. (2011). *Inside man: The discipline of modeling human ways of being*. Stanford University Press.

The cost of success or failure for proxy signals in ecological problems

Peter Nonacs

Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA

Corresponding author:
Peter Nonacs;
Email: peter.nonacs@gmail.com

doi:10.1017/S0140525X23002959, e79

Abstract

Two of John et al.'s examples of proxy failures in ecological situations are not failures: Runaway sexual selection and marsupial neonate competition. Instead, more appropriate ecological examples may be paternal genetic kin recognition and warning coloration. These differ in proxy effectiveness and failure in ways that illustrate the importance of "costs" in the evolution of ecological proxy traits.

This commentary is addressed only to John et al.'s application of their ideas to ecological situations. I contend that two of their examples of failure are, in fact, not correct. However, this does not mean that "proxy failure" could not be relevant to other ecological situations. I suggest two other possible and common situations.

The first proposed proxy failure is "runaway sexual selection," where as a hypothetical example, peahens might have a "goal" of producing the healthiest offspring, but simply "like" to mate with peacocks with the largest, brightest tails (Andersson, 1994), this despite such tails make males less healthy (e.g., imposing a cost of not being able to effectively avoid predators or skimping on investing in their immune systems). However, if one realizes that the true goal of a peahen is to produce more grandchildren than any other peahen (i.e., have the highest evolutionary fitness) then there is no proxy failure. It is unrewarding for a peahen to produce survival-maximizing, or the "healthiest", sons either through her own genotype or from a chosen healthy but less colorful male. Those sons will likely mate with few or no females and, therefore, are wasted investment from the standpoint of their mother. In contrast, a son with a large bright tail will produce many grandchildren. Hence, even in a runaway selection scenario, the proxy measure of a big, bright tail is accurate to the goal of reproductive success through the generations. The authors suggest that proxy traits are stable with runaway sexual selection because the trait and female preference are genetically linked. I suggest the reverse: This linkage evolves *because* the proxy trait accurately predicts a fitness-enhancing choice.

The second proposed proxy failure is the example of a marsupial neonate that succeeds with stronger arms in securing its mother's teat even if the initial arm investment comes at a cost for other parts of its development that eventually result in a less viable baby. This example, however, has a problem that the authors do not address – the same individual will be an agent (as a neonate) and a regulator (as a mother) in the same evolutionary conflict. How is this to be resolved; and to advantage of which part of the individual's life history? Nevertheless, if we again simply change the perspective and the goal, the proxy failure goes away. Consider that the neonate's goal is to have the highest chance of becoming a future mother relative to all other neonates in the population. It then makes no sense, evolutionarily, to gain a teat at a cost of dying before reproducing. Thus, selection on neonates should result in the optimal trade-off between investment to gain a teat and the resulting downstream probability to survive and reach motherhood. In essence, this is exactly the same goal as the neonate's mother.

I propose to the authors that there are two better ecological examples for their hypotheses. The first is kin recognition, particularly in the case of creating paternal certainty. Hypothetically, a male bird could be certain that the chick hatching from his mate's egg is indeed his offspring, if that offspring produces one or several genetically determined physical traits that could only have come from the male. These would be "greenbeards" and act as proxies for genome-level genetic kinship (Gardner & West, 2010; Nonacs, 2011) – with the result that "father" invests more heavily in providing parental care for those he accurately "knows" are his offspring. Although this would seem to be advantageous for both father and offspring, greenbeards and true genetic kin recognition appear to be quite rare in nature. The evolutionary problem is that not all hatchlings may be the offspring of the male, and there is a differential cost to revealing one's genetics. A true offspring might get some more care, but offspring revealed not to be the male's, could get killed. Thus, a non-informative system may, on average, best serve all offspring. Again, realize that a male may have survived to become a father only because his "father" failed to recognize that he was not. The overall result appears to be that males are often left relying on some environmental or behavioral proxy (e.g., "I mated with this female, so the chicks could be mine"), which is not foolproof. Therefore, how honest any paternity proxy is, may rely on whether or not females prefer genetic monogamy over nonmonogamy (Møller, 2020). Note also that the failure and lack of a genetic proxy here may not be because of deceptive signals by the agents – that is, "I am your offspring," when that is untrue – but to the advantages of not signaling at all.

The second is warning coloration. Many species that are dangerous or poisonous advertise this with bright colors and high visual contrasts. This is a proxy for potential predators to warn them that it will cost more to attack than to avoid. Therefore, the predator and the dangerous individuals may have the same shared goal: Do not engage and attack (Aubier & Sherratt, 2020). This proxy, however, is easily and often cheated by many prey species that "mimic" the danger signals and colors without actually being either dangerous or poisonous. Yet the proxy communication systems rarely collapse from the presence of cheaters, even though the proxy signals may be very inaccurate relative to the agent's true state.

The above two examples illustrate the importance of cost in the evolution (or nonevolution) of proxy signaling systems. In the first, an accurate and honest proxy signal may result in a massive loss of fitness if it is presented to the wrong individual. Here, it is the agent that risks paying a disproportionate cost. In the case of warning coloration, the costs can be extreme for the regulator, in mistaking an honest proxy signal for a dishonest one. At least as far as ecological systems are concerned, proxy success or failure will strongly depend on the costs of mistakes and which parties have to differentially pay them.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Andersson, M. (1994). *Sexual selection*. Princeton University Press.

Aubier, T. G., & Sherratt, T. N. (2020). State-dependent decision-making by predators and its consequences for mimicry. *American Naturalist*, *196*(5), E127–E144.

Gardner, A., & West, S. A. (2010). Greenbeards. *Evolution 64*, 25–38.

Møller, A. P. (2000). Male parental care, female reproductive success, and extrapair paternity. *Behavioral Ecology 11*, 161–168.

Nonacs, P. (2011). Kinship, greenbeards, and runaway social selection in the evolution of social insect cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(Suppl. 2), 10808–10815.

Proxy failure as a feature of adaptive control systems

Tuomas K. Pernu

Department of Social Sciences, University of Eastern Finland, Joensuu, Finland
http://www.tuomaspernu.london

Corresponding author:
Tuomas K. Pernu;
Email: tuomas.pernu@uef.fi

doi:10.1017/S0140525X23002844, e80

Abstract

The analysis of John et al. is lacking a fully general account of proxy failure. It is here proposed that proxy failure can be understood as a feature of all adaptive control systems. Whether proxies "fail" or "succeed" depends on the more encompassing view one can adopt for observing such systems.

I salute the ambitious goal of providing a unified account of proxy failure, and although I find the analysis important – maybe even revolutionary – I feel it is lacking one crucial element: A fully general account of the phenomenon deemed "proxy failure." I take here steps towards providing such an analysis by outlining how proxy failure can be understood as a feature of adaptive control.

Note, first, a tension inherent in the provided analysis. The aim is to give a general, unified account of a wide variety of phenomena, ranging from physiology and engineering to ecology and economics. How to speak in *one* terminology on so *many different* things? John et al. are explicit in adopting anthropomorphic terminology, and they speak in terms of "goals," "agents," and so on. This is misleading, however, as the whole project is based on the idea that proxy failure is not an intentional or social phenomenon, but something more general and primitive at its core. So, what is this general, primitive phenomenon of proxy failure?

Let us start with the very notion of "failure." To speak in terms of failure, there must be a mismatch between actual and desired outcomes. But whose desires count here? Who or what decides whether a goal has been achieved or not? And indeed: Who gets to decide what counts as a "goal"? Clearly, what counts as "goals," and "failures" to achieve them, vary depending on the perspective one adopts: One system's failure is another one's success.

The term "proxy failure," in turn, can be understood as referring to the failure of a signal to represent a system or a process one is interested in controlling. So, we have three things: A *controller system*, a *target system*, and a *signal* carrying information about the state of the target to the controller. Now, it would be wrong to equate "proxy failure" with a mere "signal failure": central to the analysis is that the target system "hacks" the proxy for its own benefit, resulting in the signal representing the desired state of the target system. In other words, proxy failure occurs when the actions of a controller system induce changes in the target system. So, we have a coupling of *adaptive systems*, linked via a feedback loop.

In control theory, *adaptive control* is a method of controlling a target under uncertain or changing conditions (Åström & Wittenmark, 2008). Under such conditions, the controller must rely on proxies, as the target system contains uncertainties. However, in cases of proxy failure we are not dealing with only one such controller, but two, as the target changes its state *in response to* the states of the controller. But who or what determines the direction of control here? This question can only be answered by looking at the interaction of these two systems from the perspective of a more encompassing system.

Consider, to make this idea of reciprocity more vivid, the case of addiction. We find it natural to say the drug use (or a behavioural model) controls the addict rather than the other way around. That there is a *failure* on part of the addict is because of us adopting the perspective of the whole person and her long-term goals. But from the perspective of the dopaminergic system there's a *success*: It's functioning as it should (in the case of "willing addicts," there might be a success even from the perspective of the whole person).

It would thus seem that proxy failure occurs when we adopt the perspective of one adaptive controller seeking to control another, and the latter adjusts its state in reaction to the former's. It seems also reasonable to assume that such phenomena occur in all real, natural processes, as such processes evolve as interactions of adaptive systems. This provides, I suggest, the basis of a general and primitive account of proxy failure.

How, then, to make the idea of adaptive control more precise? This is a difficult, open conceptual question. As a field of engineering, adaptive control is well entrenched, with its roots going back to designing systems for controlling supersonic flight (Åström & Kumar, 2014). However, as a field of science, engineering is notoriously heuristic, and it is difficult to pin down a single coherent conceptual framework for adaptive control. Moreover, it is not at all clear that the term "adaptive" in engineering can be taken as synonymous with the term "adaptive" in biology (or other relevant fields of science). The problem is that, in biology, "adaptation" refers to traits that have *emerged during the course of evolution* because of their advantage (measured in terms of fitness benefits) to a population of organisms. In engineering, in contrast, "adaptation" refers to a set of fixed traits (control parameters) of the controller, the states (parameter values) of which can vary in response to the changing conditions (as perceived by the controller).

This conceptual gap notwithstanding, it is clear that these fields of science, and the notions of adaptation they employ, have strong affinity. Indeed, it is not uncommon to see natural selection – "the blind watch maker" (Dawkins, 1986) – compared to engineering. More strikingly, in his seminal paper on evolution by natural selection, Wallace (1858) made an explicit comparison between the principle of natural selection and control theory (of the era), by noting that "this principle is exactly like that of the centrifugal governor of the steam engine" (p. 62; cf. Smith, 2004). Recently, a variety of control theoretic accounts of evolutionary processes has been advanced (e.g., Avila, Priklopil, & Lehmann, 2021; Badyaev, 2019; Cisek, 2019, 2022; Cowan et al., 2014; Lehmann, 2022).

This suggests that from an abstract point of view, we can view all organisms and evolutionary processes as hierarchies of control systems. Proxy failures can then be observed at any level of organisation, when a more general perspective on the competition of adaptive controllers has been adopted. In adaptive contexts, natural selection can be viewed as the most general controller. But natural selection cannot fall victim to proxy failure, as it does not engage in competition, but provides the very conceptual basis of it.

Financial support. This work has not been funded externally.

Competing interest. None.

References

Åström, K. J., & Kumar, P. R. (2014). Control: A perspective. *Automatica 50*, 3–43.
Åström, K. J., & Wittenmark, B. (2008). *Adaptive control*. Dover.
Avila, P., Priklopil, T., & Lehmann, L. (2021). Hamilton's rule, gradual evolution, and the optimal (feedback) control of phenotypically plastic traits. *Journal of Theoretical Biology 526*, 110602.
Badyaev, A. V. (2019). Evolutionary transitions in controls reconcile adaptation with continuity of evolution. *Seminars in Cell & Developmental Biology 88*, 36–45.
Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, and Psychophysics 81*, 2265–2287.
Cisek, P. (2022). Evolution of behavioural control from chordates to primates. *Philosophical Transactions of the Royal Society B 377*, 20200522.
Cowan, N. J., Ankarali, M. M., Dyhr, J. P., Madhav, M. S., Roth, E., Sefati, S., … Daniel, T. L. (2014). Feedback control as a framework for understanding tradeoffs in biology. *Integrative and Comparative Biology 54*, 223–237.
Dawkins, R. (1986). *The blind watchmaker*. Longman.
Lehmann, L. (2022). Hamilton's rule, the evolution of behavior rules and the wizardry of control theory. *Journal of Theoretical Biology 555*, 111282.
Smith, C. H. (2004). Wallace's unfinished business: The "Other Man" in evolutionary theory. *Complexity 10*, 25–32.
Wallace, A. R. (1858) On the tendency of varieties to depart indefinitely from the original type. *Proceedings of the Linnean Society of London 3*, 53–62.

Changing the incentive structure of social media may reduce online proxy failure and proliferation of negativity

Claire E. Robertson[a], Kareena del Rosario[a], Steve Rathje[a] and Jay J. Van Bavel[a,b,c]

[a]Department of Psychology, New York University, New York, NY, USA; [b]Center for Neural Science, New York University, New York, NY, USA and [c]Department of Strategy and Management, Norwegian School of Economics, Bergen, Norway
kareena.delrosario@nyu.edu; sr6276@nyu.edu; jay.vanbavel@nyu.edu

Corresponding author:
Claire E. Robertson;
Email: cer493@nyu.edu

doi:10.1017/S0140525X23002935, e81

Abstract

Social media takes advantage of people's predisposition to attend to threatening stimuli by promoting content in algorithms that capture attention. However, this content is often not what people expressly state they would like to see. We propose that social media companies should weigh users' expressed preferences more heavily in algorithms. We propose modest changes to user interfaces that could reduce the abundance of threatening content in the online environment.

Through millennia of evolution, the brain has developed an attentional system that preferentially orients people toward physical, emotional, and social threats (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Öhman & Mineka, 2001; Rozin & Royzman, 2001). The *proxy* for threat is, then, negative or threatening stimuli for both the individual and their group. Indeed, people attend to physical threats such as snakes or heights (Dijksterhuis & Aarts, 2003; Öhman & Mineka, 2001), social threats such as anger or out-group animosity (Fox et al., 2000; Rathje, Van Bavel, & Van Der Linden, 2021), and moral threats such as outrage or disgust (Brady, Crockett, & Van Bavel, 2020; Brady, Gantman, & Van Bavel, 2020; Hutcherson & Gross, 2011), in part because they alert people to potential harm. Our attentional system acts as a *regulator*, and its *goal* is to attend to stimuli that are relevant for people's wellbeing. *Proxy failure* occurs when people attend to false threats, even though no harm is imminent (John et al.).

Thus, attention is biased by threat proxies. It is therefore problematic that social media in particular uses attention as a proxy for what people want to see, and incorporate that into their algorithms. By using attention as a proxy for interest, social media companies motivate users to try to "hack" innate threat detection systems that drive people's attention (Brady, McLoughlin, Doan, & Crockett, 2021; Crockett, 2017). In this way, negative, divisive, and threatening content may be preferentially spread by algorithms because of lower-level proxy failure.

Critically, people do not necessarily *want* negative and divisive content in their social media feeds. Although people acknowledge that negative, false, and hateful content goes viral online, they do not want such content to go viral (Rathje, Robertson, Brady, & Van Bavel, in press). Rather, they want accurate, positive, and educational content to go viral. Proxy failure may be partially responsible for this outcome – focusing too much on social media engagement (likes, shares, etc.) as a proxy for people's preferences leads to proxy failure, because engagement metrics often promote content that people say they do not like, such as false, negative, or hostile content (Brady, Wills, Jost, Tucker, & Van Bavel, 2017; Rathje et al., 2021; Robertson et al., 2023).

Furthermore, changing incentive structures on social media toward more positive and constructive content may have downstream positive consequences. When the incentive structure of social media changes, people adjust the type of content they share. For instance, when social media platforms reward the veracity or trustworthiness of a post, people are far more likely to share true information (Globig, Holtz, & Sharot, 2023; Pretus et al., 2023). Furthermore, this can be achieved with a relatively small tweak to the design features of a social media site. Adding "trust," "distrust," or "misleading" buttons to the standard "like" and "dislike" reactions led people to become more discerning of true and false information, and more likely to share accurate information.

When attention or watch time is used as the basis for newsfeed algorithms, the algorithms may become more likely to show harmful or negative content because of peoples' attentional bias toward threats and engagement with divisive content (Milli, Carroll, Pandey, Wang, & Dragan, 2023). However, if people were given a simple way to curate their news feeds, they would be able to make their news feeds align more with their preferences. Thus, social media companies should use people's expressed preferences for what they want to see on social media (e.g., positive, nuanced, educational, or entertaining content) as a proxy for interest, rather than ambiguous metrics such as attention or engagement. This could be achieved with modest changes to existing social media platforms (Globig et al., 2023; Pretus et al., 2023).

Adding an accessible mechanism that lets users say "I don't want to see content like this" could substantially reduce people's exposure to unwanted threatening stimuli that an attentional proxy might promote. Most social media companies even already have this functionality, but it is often hidden in menus, cumbersome to activate, or the default is set to attention-based proxies. Positive results might be achieved by simply moving such a feature to the default or a more visible location in the platform design.

Overall, this type of intervention will only work if social media companies' superordinate goals include and prioritize positive user experience and societal outcomes over monetary gain. Negativity and toxicity increase grab people's attention, moral outrage, and spill into offline behaviors such as hate speech and endorsement of antidemocratic action (Brady et al., 2021; Kim, Guess, Nyhan, & Reifler, 2021; Suhay, Bello-Pardo, & Maurer, 2018). Reducing such content may lead to greater user wellbeing and a reduction in negativity, but it also reduces people's use of social media (Beknazar-Yuzbashev, Jiménez Durán, McCrosky, & Stalinski, 2022). This may explain why whistleblowers have revealed that Meta considered implementing algorithms that help downregulate content that is "bad for the world," but decided against it because it reduced time spent on the platforms by the users (Roose, Isaac, & Frenkel, 2020). If the superordinate goal is simply making money, then social media companies may continue to "hack" people's attentional systems and promote content that draws users' attention through sensationalism and threat.

Acknowledgments. The authors acknowledge the Social Identity and Morality Lab for their helpful feedback.

Financial support. This work was supported by the Templeton World Charity Foundation (J. J. V. B., TWCF-2022-30561) and the Russell Sage Foundation (J. J. V. B. and S. R., G-2110-33990).

Competing interest. None.

References

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Beknazar-Yuzbashev, G., Jiménez Durán, R., McCrosky, J., & Stalinski, M. (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*. https://doi.org/10.2139/ssrn.4307346

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010.

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*(4), 746.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.

Dijksterhuis, A., & Aarts, H. (2003). On wildebeests and humans: The preferential detection of negative stimuli. *Psychological Science*, *14*(1), 14–18.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition & Emotion*, *14*(1), 61–92.

Globig, L. K., Holtz, N., & Sharot, T. (2023). Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife*, *12*, e85767.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*(4), 719.

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Milli, S., Carroll, M., Pandey, S., Wang, Y., & Dragan, A. D. (2023) Engagement, user satisfaction, and the amplification of divisive content on social media. Preprint at arXiv https://doi.org/10.48550/arXiv.2305.16941

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483.

Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation and resistance to fact-checking. *Journal of Experimental Psychology: General*, *152*(11), 3116–3134. https://doi.org/10.1037/xge0001436

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (in press). People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(26), e2024292118.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822.

Roose, K., Isaac, M., & Frenkel, S. (2020). Facebook struggles to balance civility and growth. *The New York Times*, *24*.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, *23*(1), 95–115. https://doi.org/10.1177/1940161217740697

It's the biology, stupid! Proxy failures in economic decision making

Pier Luigi Sacco[a,b]

[a]Biobehavioral Arts & Culture for Health, Sustainability & Social Cohesion (BACH) Center, University of Chieti-Pescara Viale Pindaro, Pescara, Italy and [b]metaLAB (at) Harvard, Cambridge, MA, USA
https://www.unich.it/ugov/person/1902
pierluigi_sacco@fas.harvard.edu; https://pierluigisacco.eu

Corresponding author:
Pier Luigi Sacco;
Email: pierluigi.sacco@unich.it

doi:10.1017/S0140525X23002947, e82

Abstract

Utilitarian characterizations of economic decision making fail to capture the complex, conditional, and heterogeneous motivations underlying human behavior as shaped by the predictive, multicriterial drivers of biological regulation. Unless economic models start to acknowledge that humans have bodies and a biology with its own adaptive logic and tradeoffs, economic policies will be systematically exposed to, and systematic generators of, proxy failures.

The target article by John et al. provides a compelling framework for understanding proxy failures across biological and social systems. My contribution to the debate is specifically stressing how a key source of proxy failures in economics stems from a mismatch between the evolutionary function of biological reward systems and traditional conceptions of utility maximization. A clear illustration of this contradiction is the massive, unprecedented production and marketing of super-addictive goods that exploit vulnerabilities in dopaminergic reinforcement learning circuits, regarded by many as a legitimate (and even welfare improving) market response to individual utility-maximizing choices. This resonates with the example of neural proxy failure in Section 4.1 of the target article.

The notion of pleasure seeking as the implicit goal guiding human behavior is deeply embedded in classical utilitarian foundations of economics, from Bentham to Jevons (Sigot, 2002). However, as discussed by the authors, the dopaminergic system does not simply encode the hedonic value of pleasure. Rather, phasic dopamine provides a reward prediction error signal that enables reinforcement learning about stimuli and actions (Glimcher, 2011). Products engineered to artificially amplify dopamine release, through ingredients targeting vulnerable nodes in mesolimbic reward circuitry, lead to maladaptive overvaluation of associated representations. This parallels neural proxy failure examples involving drugs of abuse.

Crucially, the evolutionary function of dopamine is to provide a scalar approximation of expected net benefits over cumulated rewards associated with stimuli and actions (Pasquereau & Turner, 2013). But the open-ended pursuit of artificial dopamine release, enabled by modern production and marketing techniques, violates an implicit assumption of the neural valuation system evolved in resource-limited environments. The result is uncontrolled dopamine-seeking behavior at the expense of organismic regulation and health. This evolutionary mismatch, which disrupts the proxy relation between dopamine and net benefits of actions, might underlie the pursuit and supply of unprecedented varieties of super-addictive synthetic products in the modern economy.

The solution requires integrating core biological principles such as allostatic regulation into formal and conceptual models of economic decision making. The organism's implicit goal is survival and health. But a utilitarian characterization of economic incentives prescribes the maximization of pleasure-related proxies such as utility. As the target article suggests, proxy failures may be mitigated by directly incentivizing adaptive goals such as sustainability and wellbeing, rather than easily hackable proxies with ambiguous adaptive value. Overcoming such dysfunction requires reconceptualizing the motivational foundations of economics in terms of allostatic, reward-predictive principles of life regulation (Friston, 2010; Sennesh et al., 2022), not mechanistic pleasure seeking – which does not amount of course to denying the adaptive value of pleasure signals (Berridge & Kringelbach, 2015).

The mismatch between dopaminergic valuation systems evolved in a radically different environment and the modern economy leads to systematic proxy failures. But the problem runs deeper than specific hackable mechanisms. Utilitarian characterizations assume that economic decisions maximize a utility function that confuses pleasure for the organism's implicit goal. We must reconceptualize human motivation based on neurobiological principles of reward prediction and allostasis. Only then can we mitigate the proxy failure inherent in folk notions of "human desires" guiding behaviors. The solution is not correcting human decisional failures by means of paternalistic nudging interventions, but getting the biology right.

A biologically grounded approach to economic decision making may then move from the following building blocks.

Predictive regulation and allostasis: Organisms do not seek to simply maximize pleasure, but rather aim to maintain stability and homeostasis in the face of a changing environment. The brain uses interoceptive, hormonal, and sensory signals to predictively regulate physiological needs before they arise (Sterling, 2012). Allostasis refers to the continuous readjustment of optimal setpoints to minimize long-term cumulative costs arising from tradeoffs between competing demands (Schulkin & Sterling, 2019). This principle of stability through change is fundamentally at odds with notions of rational utility maximization.

Reward prediction errors, not pleasure: Dopamine neurons signal discrepancies between expected and received reward, that is, reward prediction errors (RPEs). Rather than encoding pleasure per se, phasic dopamine enables learning by stamping in stimulus associations that reduce RPEs over time (Abler, Walter, Erk, Kammerer, & Spitzer, 2006). Drugs hijack this system by artificially inducing hyper-dopaminergic states. The evolutionary function is reward-based learning, not pleasure seeking.

Multicriterion optimization: The complex dynamics of allostatic regulation imply brain activity optimizes multiple constraints, not a single utility measure. Interdependent processes spanning molecular, cellular, network, and behavioral levels interact to mediate adaptive tradeoffs (Sterling, 2020). Multicriterial decision models better capture the conditional heuristics employed by biological systems than simplifying assumptions of utility maximization (Greco, Ehrgott, & Figueira, 2016).

Motivational heterogeneity: Humans exhibit a diversity of social and cultural motives not adequately accounted for in economic models, such as meaning, autonomy, and connectedness. Intrinsically motivated behavior is guided by goal states rather than pleasure states (Ryan & Deci, 2000). Traditional notions of utility conflate hedonic experience with fundamentally distinct drivers of behavior.

In summary, utilitarian characterizations of economic decision making fail to capture the complex, conditional, and heterogeneous motivations underlying human behavior as shaped by the predictive, multicriterial drivers of biological regulation. Unless economic models start to acknowledge that humans have bodies and a biology with its own adaptive logic and tradeoffs, economic policies will be systematically exposed to, and systematic generators of, proxy failures.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, *31*(2), 790–795.

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, *86*(3), 646–664.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(S3), 15647–15654.

Greco, S., Ehrgott, M., & Figueira, J. R. (Eds.) (2016). *Multiple criteria decision analysis. State of the art surveys*. Springer.

Pasquereau, B., & Turner, R. S. (2013). Limited encoding of effort by dopamine neurons in a cost-benefit trade-off task. *Journal of Neuroscience*, *33*(19), 8288–8300.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Schulkin, J., & Sterling, P. (2019). Allostasis: A brain-centered, predictive mode of physiological regulation. *Trends in Neurosciences*, *42*(10), 740–752.

Sennesh, E., Theriault, J., Brooks, D., van de Meent, J. W., Feldman Barrett, L., & Quigley, K. S. (2022). Interoception as modeling, allostasis as control. *Biological Psychology*, *167*, 108242.

Sigot, N. (2002). Jevons's debt to Bentham: Mathematical economy, morals and psychology. *The Manchester School*, *70*(2), 262–268.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior*, *106*(1), 5–15.

Sterling, P. (2020). *What is health? Allostasis and the evolution of human design*. MIT Press.

Reductionism and proxy failure: From neuroscience to target-based drug discovery

Arash Sadri[a,b] and Sepideh Paknezhad[c]

[a]Lyceum Scientific Charity, Tehran, Iran; [b]Interdisciplinary Neuroscience Research Program (INRP), Students' Scientific Research Center, Tehran University of Medical Sciences, Tehran, Iran and [c]Department of Psychology, Sari Branch, Islamic Azad University, Sari, Iran
https://lyceum.charity
sepidehpaknezhad@outlook.com

Corresponding author:
Arash Sadri;
Email: arashsadri@lyceum.charity

doi:10.1017/S0140525X23002893, e83

Abstract

Reductionist methodologies reduce phenomena to some of their lower-level components. Researchers gradually shift their focus away from observing the actual object of study toward investigating and optimizing such lower-level proxies. Following reductionism, these proxies progressively diverge further from the original object of study. We vividly illustrate this in the evolution of target-based drug discovery from rational and phenotypic drug discovery.

We want to add reductionist methodologies to John et al.'s valuable list of the various contexts where proxy failure can be observed. Reductionist methodologies primarily focus on a lower-level component of their phenomena of interest with this presumption that understanding or manipulating that component is sufficient for understanding or manipulating their phenomenon of interest. In other words, they presume that their phenomenon of interest can be reduced to some of its lower-level components. We believe that extending the conceptualization of John et al. to reductionist methodologies and using the insights it might offer can be of great value because currently, despite considerable criticism, reductionism dominates diverse scientific fields; from psychology and neuroscience to biomedical sciences and drug discovery.

To align these methodologies with the conceptualization and terminology of John et al., researchers can be considered *regulators*, with the *goal* of understanding or manipulating some complex phenomenon manifested by their objects of study (*agent*). These objects of study "possess multiple ways of producing or expressing a proxy, which can be influenced by feedback from the regulator" (target article, sect. 3.1, para. 4). Finally, the *proxy* is a lower-level component of the phenomenon under study that the researchers prioritize. The pressure "that tends to make the proxy a worse approximation of the goal" (target article, abstract, para. 1) has been provided and amplified by the deep-seated reductionist mindset of the scientific community and the incessant technological progress that has enabled the dissection and reduction of many phenomena to their lower levels.

Consider the example of target-based drug discovery that has been the dominant paradigm of drug discovery for about four decades. Although the *goal* of researchers in the field of drug discovery (*regulators*) is to discover molecules that can suitably alter the phenotypes of humans, based on reductionist target-based drug discovery, the lower-level *proxy* of binding affinity to a target protein is prioritized. Instead of selecting and optimizing molecules based on their effects on phenotypes, candidate molecules are primarily selected and optimized based on their binding affinity to a target protein whose manipulation is supposed to counteract the disorder; for example, candidate antipsychotics are selected and optimized based on their binding affinity to specific dopamine receptors. This selection and optimization is the *regulatory feedback* mentioned by John et al. where the agents are selected based on the proxy and "which induces optimization of the proxy" (target article, sect. 3.3, para. 1).

For years, binding affinity to a target has been criticized as an overly simplistic proxy (Horrobin, 2003). Our recent analysis of the real-world efficiency of target-based drug discovery further substantiates these criticisms and reveals, based on significant evidence, the failure of this proxy (Sadri, 2023). The data reveal that, despite decades of utter dominance, only 9.4% of small-molecule-approved drugs have originated from target-based drug discovery.

Moreover, the analysis demonstrates that even this minor portion cannot be entirely attributed to target-based drug discovery, as their therapeutic effects are mediated by numerous mechanisms that are independent of the targets they have been discovered for (Sadri, 2023). This aligns perfectly with one of the factors identified by John et al. to drive proxy failure: The human body and the therapeutic effects of molecules are highly complex and there are numerous "proxy-independent actions that lead to the goal" (target article, sect. 3.3, para. 1).

Another match is between the evolution of drug discovery methodologies and the statement of John et al. that "whenever incentivization or selection is based on an imperfect proxy measure of the underlying goal, a pressure arises that tends to make the proxy a worse approximation of the goal" (target article, abstract, para. 1). Target-based drug discovery was born out of rational drug discovery (Al-Ali, 2016; Sadri, 2023). Rational drug discovery used the available scientific knowledge, from molecular biology and physiology to pathology and pharmacology, to guide and focus the random screening of substances. This approach was immensely successful and culminated in the discovery of tens of approved drugs and Nobel Prizes for several of its pioneers: Paul Ehrlich, Gertrude Elion, George Hitchings, and James Black. Assessing the activity of molecules at the protein level and their binding affinity toward specific proteins was an important proxy and source of information in rational drug discovery; however, in the end, it relied on phenotypic observations for selecting and optimizing molecules. This is evident in the discoveries of the abovementioned pioneers and many other examples (Sadri, 2023). Gradually, the proxy of binding affinity to targets got excessively prioritized to the point that it even replaced the use of phenotypic observations in selecting and optimizing molecules.

Similar cases of proxy failure can be recognized in other reductionist methodologies across different fields. For example, in neuroscience, although the *goal* is to understand the behaviors of organisms (*agents*), researchers (*regulators*) excessively prioritize various *proxies* and observations at lower levels such as neurons and neural circuits (Frangou, 2020; Gazzaniga, 2010; Krakauer et al., 2017; Parker, 2022; Uttal, 2003). This conceptualization can be extended to the reductionist methodologies that are currently dominant across various fields, including molecular biology (Lazebnik, 2002) and medical sciences (Ahn et al., 2006).

However, it must be noted that although John et al.'s conceptualization perfectly fits the reductionist methodology of target-based drug discovery, extending it to other reductionist methodologies may require a more general account of proxy failure. The challenge is generalizing to these methodologies the key concept of the pressure "that tends to make the proxy a worse approximation of the goal" (target article, abstract, para. 1). Alternatively, it can be proposed that another approach toward generalizing the conceptualization of John et al. to reductionist methodologies might mitigate this challenge. For example, reductionism and the reductionist methodologies themselves can be considered as the regulator, the researchers as the agents, investigating the object of study as the goal, and the lower-level observations as the proxies.

Anyhow, we believe that the concept of proxy failure is a valuable asset for recognizing and addressing the limitations of reductionism, which is in absolute reign across diverse fields. Lest these limitations are addressed, huge resources would be expended on research with marginal contact with the real world.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Ahn, A. C., Tewari, M., Poon, C. S., & Phillips, R. S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Medicine*, *3*(6), e208. https://doi.org/10.1371/journal.pmed.0030208

Al-Ali, H. (2016). The evolution of drug discovery: From phenotypes to targets, and back. *MedChemComm*, *7*(5), 788–798. https://doi.org/10.1039/C6MD00129G

Frangou, S. (2020). The enduring allure of explanatory reductionism in schizophrenia. *Biological Psychiatry*, *87*(8), 690–691. https://doi.org/10.1016/j.biopsych.2020.01.008

Gazzaniga, M. S. (2010). Neuroscience and the correct level of explanation for understanding mind. *Trends in Cognitive Sciences*, *14*(7), 291–292. https://doi.org/10.1016/j.tics.2010.04.005

Horrobin, D. F. (2003). Modern biomedical research: An internally self-consistent universe with little contact with medical reality? *Nature Reviews Drug Discovery*, *2*(2), 151–154. https://doi.org/10.1038/nrd1012

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Lazebnik, Y. (2002). Can a biologist fix a radio? – Or, what I learned while studying apoptosis. *Cancer Cell*, *2*(3), 179–182. https://doi.org/10.1016/s1535-6108(02)00133-2

Parker, D. (2022). Neurobiological reduction: From cellular explanations of behavior to interventions. *Frontiers in Psychology*, *13*, 987101. https://doi.org/10.3389/fpsyg.2022.987101

Sadri, A. (2023). Is target-based drug discovery efficient? Discovery and "off-target" mechanisms of all drugs. *Journal of Medicinal Chemistry*, *66*(18), 12651–12677. https://doi.org/10.1021/acs.jmedchem.2c01737

Uttal, W. R. (2003). *The new phrenology: The limits of localizing cognitive processes in the brain.* MIT Press.

Navigating proxy failures in education: Learning from human and animal play

Robin Samuelsson

Department of Scandinavian Languages, Uppsala University, Uppsala, Sweden
robin.o.samuelsson@nordiska.uu.se

doi:10.1017/S0140525X23002972, e84

Abstract

The notion of proxy failure provides considerable insight into educational processes, and in childhood education has the potential to elucidate known problems stemming from the early implementation of overly regulated educational regimes. This commentary expands on play and how its relation to learning provides a useful perspective on how activities based on non-goal-oriented interactions can lead to desired outcomes.

John et al. provided an underlying theoretical framework explaining the widely recognized phenomenon of why a measure ceases to be good when it becomes the target for what it aims to measure, capturing this phenomenon, termed *proxy failure*, in a range of domains. Associated problems have been widely debated in education, often under the banner of Campbell's law, and popularized in the notion of "training for the test," where standardized tests only measure a subset of the properties they aim to measure (Koretz, 2008), leading to a variety of problems such as a too-narrow focus on end outcomes (i.e., the measure of learning chosen) rather than evaluating the activities, interactions, and environments that lead to those outcomes.

Paradoxically, relying solely on educational goals can lead to worse learning. This is identified in John et al.'s proxy failure concept. When proxies are adopted to steer the direction of education, even well-intentioned educational efforts eventually miss their goal. In childhood education, the counterintuitive results that academically goal-oriented educational programs do not produce better academic outcomes (e.g., Durkin, Lipsey, Farran, & Wiesen, 2022) might be explained by how the focused training children receive provides a too-narrow proxy for the complex phenomenon of learning. Problems with such programs have led scholars to turn to play as a cornerstone of learning and a serious educational alternative for young children (Nesbitt, Blinkoff, Golinkoff, & Hirsh-Pasek, 2023). It can also hold other important implications, as play and its relationship with learning can provide insights into proxy failure mechanisms and provide alternatives for education and beyond.

Play as a phenomenon has long been a puzzle for scientists as it superficially looks like a nongoal-oriented and perhaps purposeless activity (Pellegrini, 2009). It is more puzzling still that play is adaptive for young organisms across species (Burghardt, 2005). Crucially, because the focus of play is not on any external goal, engaging in play and related exploratory behavior trains the organism for uncertain outcomes in unstable environments (Bjorklund, 2022). For humans, this takes a cultural turn as children play with and socialize around the tools and technologies in their environment (Samuelsson, Price, & Jewitt, 2022), leading to culturally appropriate learning potential as a by-product of this engagement (Samuelsson, 2023). Thus, play provides a valuable perspective on how less-regulated activities can lead to learning outcomes.

Play provides a peculiar window into the learning process that differs radically from many formal educational environments, for in play, learning is not a predetermined goal but follows as a by-product of engagement in an activity. Additionally, play is often a joyful experience for children (Weisberg, Hirsh-Pasek, Golinkoff, Kittredge, & Klahr, 2016). We can learn from studies of play, as they show how a different type of activity not associated with external goals also leads to learning results that are on par with or even better (e.g., Toub et al., 2018) than those of a goal-oriented learning process.

This is also telling regarding proxy failures related to education. If education is based on narrow learning goals, it risks the proxy treadmilling effect identified by John et al., where teachers can too narrowly train for the test, or children find ways to "hack" the regulatory system. By contrast, playful learning is less concerned with external learning goals. It is centrally characterized by focused engagement in joyful activities and, if used correctly, engages curious children at the height of their abilities (Chu & Schulz, 2020). In "hacking" or "gaming," the regulatory system becomes pointless, as it is not the goal of the activity (e.g., there is no question whether something will be on the test). However, play may involve practice in iterative engagement, which is typical of play. From this engagement, play becomes "training for the unexpected" (Spinka, Newberry, & Bekoff, 2001, p. 143), a widely observed function of play. There is something to be learned from play's relation to learning because of its nongoal-oriented nature precisely because there is no proxy associated with a learning goal that leads to learning.

Playful learning has been widely adopted in childhood educational programs such as guided play activities (Weisberg et al., 2016). Overall, these initiatives can tailor the positive aspects of play in educational environments that are well adapted to childhood education. When using play for learning outcomes, however, one must be careful. Paradoxically, educational programs based on play-based learning risk are becoming examples of proxy failures themselves if they are abundantly goal-driven, reducing the robustness of playful engagement for narrower and more easily measured learning outcomes (e.g., the legibility problem). Here, John et al. warn us about the mechanistic properties of proxy failure, meaning that even if actors in an educational system are well-intentioned, proxies move toward failure. Educational systems wishing to use play as part of their learning program should learn from play that the loosening of educational proxies can lead to learning and consider how the inherent looseness of playful activities yields robust learning.

Understanding the problems instilled by creating measurable proxies for complex learning processes entails a considerable rethinking of how we conduct education. It is, however, a key lesson for not only creating educational settings that avoid the narrowing associated with failed educational proxies, but also for other areas wishing to navigate proxy failures. As play shows, engaging in activities with internally driven goals can lead to robust outcomes, such as long-term, meaningful learning. This can hold implications beyond the early childhood classroom and make us more humble in creating proxies for goal-driven processes, considering the often-nested hierarchies imposed when creating a target measure. Potential regulatory agents can benefit from an awareness of proxy failure mechanisms, here meaning all actors from the education providers of play-based learning to the teachers who engage in guided play interactions with children. This is one example of how John et al.'s contribution to proxy failure constitutes fundamental work for understanding educational processes and can be a crucial instrument for creating educational futures of genuine quality.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Bjorklund, D. F. (2022). Children's evolved learning abilities and their implications for education. *Educational Psychology Review*, *34*, 2243–2273. doi:10.1007/s10648-022-09688-z

Burghardt, G. M. (2005). *The genesis of animal play: Testing the limits*. MIT Press.

Chu, J., & Schulz, L. (2020). Play, curiosity, and cognition. *Annual Review of Developmental Psychology 2*, 317–343. doi:10.1146/annurev-devpsych-070120-014806

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology*, *58*(3), 470–484. doi:10.1037/dev0001301

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Nesbitt, K. T., Blinkoff, E., Golinkoff, R. M., & Hirsh-Pasek, K. (2023). Making schools work: An equation for active playful learning. *Theory Into Practice*, *62*(2), 141–154. doi:10.1080/00405841.2023.2202136

Pellegrini, A. D. (2009). *The role of play in human development*. Oxford University Press.

Samuelsson, R. (2023). Leveraging play for learning and development: Incorporating cultural-evolutionary insights into early educational practices. Mind, *Brain, and Education*, *17*(2), 75–85. doi:10.1111/mbe.12347

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Samuelsson, R., Price, S., & Jewitt, C. (2022). How young children's play is shaped through common iPad applications: A study of 2 and 4–5 year-olds. *Learning, Media and Technology*, 1–19. doi:10.1080/17439884.2022.2141252

Spinka, M., Newberry, R. C., & Bekoff, M. (2001). Mammalian play: Training for the unexpected. *The Quarterly Review of Biology*, *76*(2), 141–168. doi:10.1086/393866

Toub, T. S., Hassinger-Das, B., Nesbitt, K. T., Ilgaz, H., Weisberg, D. S., Hirsh-Pasek, K., … Dickinson, D. K. (2018). The language of play: Developing preschool vocabulary through play following shared book-reading. *Early Childhood Research Quarterly*, *45*, 1–17. doi:10.1016/j.ecresq.2018.01.010

Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., Kittredge, A. K., & Klahr, D. (2016). Guided play: Principles and practices. *Current Directions in Psychological Science*, *25*(3), 177–182. doi:10.1177/0963721416645512

Proxy failure in social policies as one of the main causes of persistent sexism and racism

Konrad Szocik[a,b]

[a]Department of Social Sciences, University of Information Technology and Management in Rzeszow, Rzeszow, Poland and [b]Interdisciplinary Center for Bioethics, Institution for Social and Policy Studies, Yale University, New Haven, CT, USA

Corresponding author:
Konrad Szocik;
Email: konrad.szocik@yale.edu

doi:10.1017/S0140525X23002996, e85

Abstract

One example of proxy failure is current antisexist and antiracist policies. One of the most popular proxy in them is the number of representatives of marginalized groups – women and nonwhite people – in power structures. Here I show that such measures do not lead to combating sexism and racism, which flourish despite their application.

In this commentary, the theory of proxy failure is extended to social systems. John et al. discuss three examples from neuroscience, economics, and ecology. I propose to apply this theory in the field of antisexist and antiracist policies. The minimal goal of fighting sexism and racism is to guarantee equal rights and opportunities for women (sexism) and nonwhites (racism). It is important to consider what is the regulator in both cases. The regulator is both the government and state institutions, as well as private, civic institutions, and can also be individuals. The problem, therefore, is the dispersion of responsibility, as well as conflicts of interest at all levels of potential regulators. Similarly complex is the structure of agents, which can be simultaneously many institutions as well as individuals. What is least ambiguous is the proxy. In both social policies, the proxy is usually the percentage of those excluded in a given power structure, probably the most common and easiest to measure. Here I show that antisexist and antiracist policies are examples of proxy failure. They do not lead to the actual goal of equality. They are also in line with the ideal of liberal feminism, whose goal – and perhaps a side effect caused by proxy failure – is to preserve the unequal structure by focusing on the aforementioned proxy, rather than changing the structure in a more equal direction.

Racism is a very strong system of domination, resistant to antiracist discourse (van Dijk, 2021). The goal of antiracism is to combat the inequality caused by racism. The regulator of antiracism policy includes all levels of society. Moreover, as history shows, antiracism is basically impossible without grassroots action by the disadvantaged and excluded themselves. This results in a vicious circle paradox. This is because we expect activity and initiative from those whose initiative and activity have been and are being restricted. Another problem of antiracist policies is proxy. The percentage of candidates hired does not address other problems that are difficult to measure or not measurable at all, such as white privilege, racism in education and justice system, color blindness, or racialization of poverty. Increasing the number of representatives of other races will not remove racism. This is because racism is based on the ideology of white supremacy, as well as being acquired through learning (Brookfield, 2019a). It is worth remembering the limited trust of people of color in whites, where any antiracist effort by whites – including parity policies – can only be perceived as another subtle attempt to mask white supremacy, which in practice cannot shake white dominance (Brookfield, 2019b). Consequently, antiracism policies should include a number of multilevel proxies, both for institutions and individuals (Berman & Paradies, 2010). Focusing on the increased numbers of racial minorities as a major proxy paradoxically can reinforce racist sentiment in the form of backlash, including in academia, which should be particularly aware of knowledge about the roots of racism and discrimination. The presence of nonwhite academics may be desirable only because of the institution's desire to appear antiracist, and not necessarily to be so in practice (Joseph-Salisbury & Connelly, 2021; Thobani, 2021). This is evidenced by the marginal treatment in the academic world of humanities and social sciences research in both gender and feminism, and especially those representing the achievements of non-Western cultures (see, as an example, the virtual absence of African or South American philosophy and the dominance of the classical canon of European philosophy presented as universal philosophy, "Philosophy," as well as the absence or minimal presence of African studies programs in most curricula).

Similarly ineffective are antisexist policies whose proxy is to increase the percentage of women in power structures. These are only apparent measures that do not eliminate sexism, but instead allow a small group of women to participate in the prosperity of the patriarchy traditionally held by men (Arruzza, Bhattacharya, & Fraser, 2019). Among traditionally marginalized groups, only able-bodied white women usually have a chance to be included in power structures, while racialized minority women are the most marginalized. Thinking about antisexist policies in terms of numerically measurable equality ignores the global perspective of millions of women around the world, whose lives will not be changed by any proxy measured by the number of women in power structures. It is about the unfair effects of climate change, which affect girls and women often more severely than men. This is especially true in those situations where women, traditionally associated with caring for the household and taking care of the family, are required to procure and prepare food, obtain increasingly difficult to access water, and are often forced to abandon their education (Atrey, 2023). Finally, women around the world are experiencing misogynist and sexist abuse and harassment, which is being intensified through the Internet (Ging & Siapera, 2019). Women still face various obstacles in many parts of the world, including the seemingly most feminized Western societies, simply because they are women, usually centered around their reproductive biology, such as difficult access to abortion, conscience clauses on the part of doctors and pharmacists, or employer interest in reproductive plans and the gender pay gap. All of these sexist and at least partially misogynist phenomena are occurring today, some of them intensifying, such as restrictions on reproductive rights in countries with previously more liberal abortion legislation, such as the USA and Poland.

These phenomena are occurring despite the continuous increase in the number of women in various power structures and social spheres. Thus, this shows that the dynamic of sexism is independent of the aforementioned proxy of antisexist policies. The identical situation applies to the fight against racism and the ineffectiveness of antiracist policies. This therefore suggests that the current proxy for these policies is flawed. It is necessary to increase the participation of both representatives of other races and women in power structures for a variety of reasons, but it is important to keep in mind that these measures do not combat either racism or sexism, and require framing new proxies.

Financial support. This work was supported by the National Science Centre, Poland (UMO 2021/41/B/HS1/00223).

Competing interest. None.

References

Arruzza, C., Bhattacharya, T., & Fraser, N. (2019). *Feminism for the 99 percent. A manifesto*. Verso.

Atrey, S. (2023). The inequality of climate change and the difference it makes. In C. Albertyn, M. Campbell, H. Alviar García, S. Fredman, & M. Rodriguez de Assis Machado (Eds.), *Feminist frontiers in climate justice: Gender equality, climate change and rights* (pp. 17–39). Edward Elgar.

Berman, G., & Paradies, Y. (2010). Racism, disadvantage and multiculturalism: Towards effective anti-racist praxis. *Ethnic and Racial Studies*, *33*(2), 214–232.

Brookfield, S. D. (2019a). The dynamics of teaching race. In S. D. Brookfield and associates, *Teaching race: How to help students unmask and challenge racism* (pp. 1–17). Jossey-Bass.

Brookfield, S. D. (2019b). Avoiding traps and misconceptions in teaching race. In S. D. Brookfield and associates, *Teaching race: How to help students unmask and challenge racism* (pp. 293–309). Jossey-Bass.

Ging, D., & Siapera, E. (2019). Introduction. In D. Ging & E. Siapera (Eds.), *Gender hate online* (pp. 1–17). Palgrave Macmillan.

Joseph-Salisbury, R., & Connelly, L. (2021). *Anti-racist scholar-activism*. Manchester University Press.

Thobani, S. (Ed.) (2021). *Coloniality and racial (in)justice in the university: Counting for nothing?* University of Toronto Press.

van Dijk, T. A. (2021). *Antiracist discourse: Theory and history of a micromovement*. Cambridge University Press.

Genies, lawyers, and smart-asses: Extending proxy failures to intentional misunderstandings

Tomer D. Ullman[a] and Sophie Bridgers[a,b]

[a]Department of Psychology, Harvard University, Cambridge, MA, USA and [b]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
www.tomerullman.org
secb@mit.edu

Corresponding author:
Tomer D. Ullman;
Email: tullman@fas.harvard.edu

doi:10.1017/S0140525X23002820, e86

Abstract

We propose that the logic of a genie – an agent that exploits an ambiguous request to intentionally misunderstand a stated goal – underlies a common and consequential phenomenon, well within what is currently called proxy failures. We argue that such intentional misunderstandings are not covered by the current proposed framework for proxy failures, and suggest to expand it.

Making your way through busy market stalls, you chance upon an antique lamp. As you brush the dust off to inspect it, a genie springs out in a cloud of colorful smoke.

"One wish, no more, no less," says the genie.

"Make me rich!" you reply, and immediately alarm bells go off in your head. Your mind floods with tragic tales of people who got what they asked for.

"Don't kill my parents so I inherit their money," you hasten to add. "Actually, don't kill anyone. Also I don't want stolen mafia money. Or any kind of stolen money. Make that 'no crimes'. And don't make it that I turn things into gold, I just want money. Real money. And don't make everyone poor so I'm rich by comparison, and…" you trail off, thinking it through.

Eventually, you place the lamp down carefully and say, "You know what? Forget I even asked."

Compared to the examples of proxy failure in John et al., our genie example seems fanciful. But, we propose the logic of the genie – an agent that exploits ambiguous requests to intentionally misunderstand the stated goal – underlies a phenomenon that is (1) well within "proxy failures," and (2) common and consequential, but (3) not covered by the current framework of John et al. Our point is not that John et al.'s framework is wrong; we find it both enlightening and useful. Rather, we suggest that many important situations that seem to fall under the notion of proxy failure require expanding their framework. Our argument is not a "no," but a "yes, and."

To start, the dynamics of intentionally misunderstanding requests follow the logic of many of the examples John et al. use when introducing the problem of interest. A tenant asked by their landlord to "do some weeding," who then pulls out three weeds and calls it a day, is acting in line with the terms used by John et al.: A regulator (landlord) with a goal (cleared yard) conveys the goal to an agent (tenant), but uses language that doesn't match the goal directly, after which the agent engages in "hacking" or "gaming."

Intentional misunderstandings are common and important. They show up in history (Scott, 1985), fables and art (Da Silva & Tehrani, 2016; Uther, 2004), childhood (Opie & Opie, 2001), and interpersonal conflict (Bridgers, Taliaferro, Parece, Schulz, & Ullman, 2023). Such letter-versus-spirit of the law concerns are also often discussed in the legal realm (Hannikainen et al., 2022; Isenbergh, 1982; Katz, 2010). But such concerns are not with scalar proxies standing in for a true but unknowable goal. Consider a lawyerly child watching videos on their tablet who is told, "time to put the tablet down," and proceeds to place the tablet on a table, only to keep watching their videos. Such a child is not optimizing a scalar reward conveyed by a parent who cannot convey some complex goal and resorts to a proxy. The parent was being quite clear, and the child was being quite a smart-ass.

If we accept the above, then intentional misunderstandings pose a challenge for the framework of John et al. The framework supposes that a regulator has a difficult-to-convey goal, and instead gives an agent a different goal. But in many current models of human communication, concepts (including goals) are hidden variables, conveyed indirectly through ambiguous utterances (Goodman & Frank, 2016). This is true for any goal, including proxies. The process of recovering meaning from ambiguous utterances is usually so transparent that people don't even notice it unless it breaks down, for example, when hijacked intentionally. To see this, take the Hanoi rats (please): The original goal of killing all the rats is unobserved, but can be easily recovered from the utterance "kill all the rats." The utterance "bring me rat tails" is not a proxy goal, it is an utterance, which can be used to derive the original goal, and is likely understood by people to mean the original goal. True, the proxy utterance can be intentionally misunderstood, but so can the original utterance – there is nothing inherently special about proxy utterances in terms of clarity from the standpoint of a theory of communication, and likewise there is nothing inherently special about a proxy goal in terms of observability.

Our differing analysis for intentional misunderstandings is important for at least two reasons: First, it moves the focus away from illegibility and prediction, especially in communicating goals to machines. People who supposedly convey a "proxy goal" to a machine do not experience failure because they can't predict all the ways in which their proxy goal might have unintended consequences. Rather, they experience failure because they didn't even realize they were conveying a different goal in the first place (many of the examples in Krakovna, 2020, are like this). An engineer evaluating a loss function infers the goal behind it, because that is how human communication works, but most machines currently aren't built to run an inference process from a loss function to an intended goal. This brings us to a second reason the differing analysis is important: It suggests currently unexplored remedies, at least for some cases. Telling a genie (or child, or lawyer) a goal, and then tacking on a long list of caveats will not stop them if they are determined to misunderstand: Every caveat is an opportunity for another loophole (in line with the "proxy treadmill"). By contrast, highlighting common-ground in order to specifically rule out loopholes is useful, for example, telling a child "can you do your homework?" and following it with "you know what I mean" to avoid the tired "yes, I *can*." Highlighting common-ground would not be useful for a machine that is optimizing a given loss function rather than engaging in human-like communication.

And what of our genie? They forgot you even asked, just like you wanted. And so, they offer you one wish. No more, no less.

Acknowledgment. We acknowledge GU for unintentionally providing some of the examples of intentional misunderstandings.

Financial support. T. D. U. is supported by a Jacobs Foundation Fellowship, and this commentary makes reference to work supported by an MIT Simons Center for the Social Brain Postdoctoral Fellowship (S. B.), and an NSF Science of Learning and Augmented Intelligence Grant 2118103 (T. D. U. and S. B.).

Competing interest. None.

References

Bridgers, S. E. C., Taliaferro, M., Parece, K., Schulz, L., & Ullman, T.. (2023). Loopholes: A window into value alignment and the communication of meaning. *PsyArxiv*.

Da Silva, S. G., Tehrani, J. J. (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society open science*, *3*(1), 150645.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Hannikainen, I. R., Tobia, K. P., de Almeida, G. D. F., Struchiner, N., Kneer, M., Bystranowski, P., … Żuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), e2206531119.

Isenbergh, J. (1982). Musings on form and substance in taxation. HeinOnline.

Katz, L. (2010). A theory of loopholes. *The Journal of Legal Studies*, *39*(1), 1–31.

Krakovna, V. (2020). Specification gaming examples in AI – Master list. http://bit.ly/kravokna_examples_list (accessed: 2020-12-28).

Opie, I. A., & Opie, P. (2001). *The lore and language of schoolchildren*. New York Review of Books.

Scott, J. C. (1985). *Weapons of the weak: Everyday forms of peasant resistance*. Yale University Press.

Uther, H.-J. (2004). *The types of international folktales – A classification and bibliography*. Suomalainen Tiedeakatemia Academia Scientiarum Fennica Exchange Centre.

Proxy failure in academia: More than just another example

Milind Watve

Independent Researcher, Pune, India
https://milindwatve.in/

Corresponding author:
Milind Watve;
Email: milind.watve@gmail.com

doi:10.1017/S0140525X23002984, e87

Abstract

Proxy failure in academia has progressed much ahead of what John et al. describe. We see advanced phenomena such as proxy complimentarity in which different players push each others' proxy failures; proxy exploitation in which external agents exploit players' proxies and predatory proxies that devour the goal itself. Academics need to avoid proxy failures by designing behaviorally sound systems.

John et al. make a remarkable articulation of the phenomenon of proxy failure although many components of it have been already well known. However, I am surprised that they do not adequately cover proxy failure in academia apart from a passing mention. I feel it is necessary to deal elaborately with proxy failure in academia for two reasons. One is that proxy failure has reached unprecedented and unparalleled levels in academia leading to bad incentives (Rammohan & Rela, 2021; Roy & Edwards, 2023; Stephan, 2012) so that we can easily identify phenomena much ahead of what John et al. describe. Apart from the three consequences of proxy failure, namely proxy treadmill, proxy cascade, and proxy appropriation (John et al.) at least three more levels are observed in academia that might be difficult to find in other fields.

Proxy complimentarity: In this, more than one type of actors benefit in different ways from a proxy and therefore they reinforce each others' dependence on the proxy resulting in a rapidly deteriorating vicious cycle. Because prestige of a journal is decided by the citations its papers receive and the impressiveness of the curriculum vitae (CV) of a researcher is decided by the impact factors of the journals, the two selfish motives complement each other in citation manipulation (Fong & Wilhite, 2017). The reviewers also may pressurize the authors to cite their papers (Teixeira da Silva, 2017) and the authors may agree in return of paper acceptance (Chawla, 2023). Institutions and funding agencies are benefited because bibliographic indices lead to a pretense of evaluation saving the cost of in-depth reading of a candidate's research (Watve, 2023). Such mutually convenient positive feedback cycles can potentially drive rapid deterioration of the goal.

Proxy exploitation: This is a phenomenon in which apart from the agents in the game optimizing their own cost benefits, a party external to the field achieves selfish benefits by exploiting proxies in the field. In academic publishing profit-making publishers of journals thrive almost entirely on journal prestige as the proxy. Editorial boards appear to strive more for journal prestige than the soundness and transparency of science (Abbott, 2023). More prestigious journals often have higher author charges (Triggle, MacDonald, Triggle, & Grierson, 2022) and make larger profits with little contribution to the original goals.

Predatory proxy: This might be the most advanced and disastrous form of proxy failure where the proxy devours the goal itself. The process of proxy appropriation where the higher-level goal does a corrective hacking of lower-level proxies (John et al.) does not work in epistemology because the goal itself is difficult to define. In business, the higher-level player directly monitors the goal of profit making and accordingly controls proxies at lower level. This does not work in academia because the higher-level organizations themselves do not have an objective perspective of the goal. As a result not only the proxies are used to evaluate individual researcher, but also they might often be confused with the progress of science itself.

In many fields of science highly complex work involving huge amounts of data and sophisticated methods of analysis are being published in prestigious journals adding little real insights to the field. For example in diseases such as type 2 diabetes, in spite of huge amount of research being published and funds being allocated, there is no success in preventing, curing, reducing mortality (ACCORD study group et al., 2008; Lee et al., 2021; Ojha, Vidwans, & Watve, 2022; The NICE-SUGAR Study Investigators, 2009) or even addressing the accumulating anomalies (Diwekar-Joshi & Watve, 2020) in the underlying theory. Nevertheless large numbers of papers continue to get published, huge amount of funding is allotted which by itself is viewed as "success" in the field. Failure of achieving the goal is not a crime in science, but quite often the failure is disguised as "success" and researchers receive life-time "achievement" awards. No scientist receiving any such awards appears to have admitted that they have actually failed to "achieve" the real goals. Efforts of a researcher, failed by this definition, should still be appreciated but it should not be called "success" or "achievement" just based on proxies. The worst outcome of proxy failure in academia is the failure to identify research failure as a failure. Many other fields including theoretical particle physics or string theory have received similar criticism (Castelvecchi & Nature magazine, 2015; Hossenfelder, 2022).

The three outcomes of proxy failure might be the reason why the creativity and disruptive content in published science is deteriorating even when it is measured by proxy itself (Chu & Evans, 2021; Park, Leahey, & Funk, 2023). Simultaneously the frequency of research misconduct, data fabrication, reproducibility crisis, paper mills, predatory journals, citation manipulations, peer-review biases, and paper retractions are alarming (Fanelli, 2009; Ioannidis, 2005; Smith, 2006; Xie, Wang, & Kong, 2021) and appear to be rising (de Vrieze, 2021).

Interestingly, many researchers working on aspects of human behavior, system design, or behavior-based policy that are potentially relevant to academia avoid talking about the academic systems (Chater & Loewenstein, 2022; Cushman, 2019). The academic system is the nearest, most accessible and most relevant system to be studied. This is the second important reason why studying proxy failure in academia needs to be prioritized. However, research addressing behavioral aspects of academia is scanty and fragmentary (Chapman et al. 2019; Clark & Winegard, 2020; Watve, 2017, 2023) and not yet even close to addressing the haunting questions at a system design or s-frame level (Chater & Loewenstein, 2022). Unless researchers address the issues in their own field and come out with sound solutions; unless they design their own systems that are behaviorally sound and little prone to proxy failure; unless they are able to minimize flaws and make the system work smoothly toward the goals, why should other fields follow their advice to redesign their systems? The natural first reaction of a citizen like me working outside mainstream academia is "Academics, mend your house first!!"

Acknowledgments. I thank Preetee O. for useful discussions in a conceptualization stage.

Financial support. There was no funding for this work.

Competing interest. None.

References

Abbott, A. (2023). Strife at eLife: Inside a journal's quest to upend science publishing. *Nature*, *615*(7954), 780–781. https://doi.org/10.1038/d41586-023-00831-6

ACCORD Study Group, Gerstein, H. C., Miller, M. E., Byington, R. P., Goff, D. C., Jr., Bigger, J. T., … Friedewald, W. T. (2008). Effects of intensive glucose lowering in type 2 diabetes. *The New England Journal of Medicine*, *358*(24), 2545–2559. https://doi.org/10.1056/NEJMoa0802743

Castelvecchi D and Nature magazine. (2015). Is string theory science? *Scientific American*, December 23, 2015.

Chapman, C. A., Bicca-Marques, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., … Chr Stenseth, N. (2019). Games academics play and their consequences: How authorship, *h*-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *286*(1916), 20192047. https://doi.org/10.1098/rspb.2019.2047

Chater, N., & Loewenstein, G. (2022). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, *46*, E147. doi:10.1017/S0140525X22002023

Chawla D. S. (2023). Researchers who agree to manipulate citations are more likely to get their papers published. *Nature*, Advance online publication. https://doi.org/10.1038/d41586-023-01532-w

Chu, J. S. G., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences of the United States of America*, *118* (41), e2021636118. https://doi.org/10.1073/pnas.2021636118

Clark, C. J., & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry*, *31*(1), 1–22. doi:10.1080/1047840X.2020.1721233

Cushman, F. (2019). Rationalization is rational. *The Behavioral and Brain Sciences*, *43*, e28. https://doi.org/10.1017/S0140525X19001730

de Vrieze J. (2021). Landmark research integrity survey finds questionable practices are surprisingly common. Science Insider, July 7th, 2021 https://www.science.org/content/article/landmark-research-integrity-survey-finds-questionable-practices-are-surprisingly-common

Diwekar-Joshi, M., & Watve, M. (2020). Driver versus navigator causation in biology: The case of insulin and fasting glucose. *PeerJ*, *8*, e10396. https://doi.org/10.7717/peerj.10396

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS ONE*, *12*(12), e0187394. https://doi.org/10.1371/journal.pone.0187394

Hossenfelder, S. (2022). No one in physics dares say so, but the race to invent new particles is pointless. *The Guardian*, September 26th, 2022. https://www.theguardian.com/commentisfree/2022/sep/26/physics-particles-physicists

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2* (8), e124. https://doi.org/10.1371/journal.pmed.0020124

Lee, C. G., Heckman-Stoddard, B., Dabelea, D., Gadde, K. M., Ehrmann, D., Ford, L., … Diabetes Prevention Program Research Group. (2021). Effect of metformin and lifestyle interventions on mortality in the diabetes prevention program and diabetes prevention program outcomes study. *Diabetes Care*, *44*(12), 2775–2782. https://doi.org/10.2337/dc21-1046

Ojha, A., Vidwans, H. & Watve, M. (2022). Does sugar control arrest complications in type 2 diabetes? Examining rigor in statistical and causal inference in clinical trials. *Medrxiv* https://doi.org/10.1101/2022.08.02.22278347

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*, 138–144. https://doi.org/10.1038/s41586-022-05543-x

Rammohan, A., & Rela, M. (2021). Ranking list for scientists: From heightening the rat-race to fraying the scientific temper. *Journal of Postgraduate Medicine*, *67*(2), 91–92. doi:10.4103/jpgm.JPGM_112_21

Roy, S., & Edwards, M. A. (2023). NSF fellows' perceptions about incentives, research misconduct, and scientific integrity in STEM academia. *Science Reports*, *13*, 5701. https://doi.org/10.1038/s41598-023-32445-3

Smith, R. (2006). Research misconduct: The poisoning of the well. *Journal of the Royal Society of Medicine*, *99*(5), 232–237. doi:10.1177/014107680609900514

Stephan, P. (2012). Perverse incentives. *Nature*, *484*, 29–31. https://doi.org/10.1038/484029a

Teixeira da Silva, J. A. (2017). The ethics of peer and editorial requests for self-citation of their work and journal. *Medical Journal Armed Forces India*, *73*(2), 181–183. doi:10.1016/j.mjafi.2016.11.008

The NICE-SUGAR Study Investigators. (2009). Intensive versus conventional glucose control in critically ill patients. *New England Journal of Medicine*, *360*(13), 1283–1297. https://doi.org/10.1056/NEJMoa0810625

Triggle, C. R., MacDonald, R., Triggle, D. J., & Grierson, D. (2022). Requiem for impact factors and high publication charges. *Accountability in Research*, *29*(3), 133–164. doi:10.1080/08989621.2021.1909481

Watve, M. (2017). Social behavioural epistemology and scientific publishing. *Journal of Genetics*, *96*, 525–533.

Watve, M. (2023). Behavioral optimization in scientific publishing. *Qeios*, https://doi.org/10.32388/8W10ND.3

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, *27*, 41. https://doi.org/10.1007/s11948-021-00314-9

Authors' Response

Teleonomy, legibility, and diversity: Do we need more "proxynomics"?

Oliver Braganza[a,b]*, Yohan J. John[c], Leigh Caldwell[d] and Dakota E. McCoy[e,f,g]

[a]Institute for Experimental Epileptology and Cognition Research, University of Bonn, Bonn, Germany; [b]Institute for Socioeconomics, University of Duisburg-Essen, Duisburg, Germany; [c]Neural Systems Laboratory, Department of Health and Rehabilitation Sciences, Boston University, Boston; [d]Irrational Agency, London, UK; [e]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA; [f]Hopkins Marine Station, Stanford University, Pacific Grove, CA, USA and [g]Department of Biology, Duke University, Durham, NC, USA

oliver.braganza@ukbonn.de
yohan@bu.edu
leigh@irrationalagency.com
mccoy6@stanford.edu

*Corresponding author.

doi:10.1017/S0140525X24000050, e88

Abstract

When a measure becomes a target, it ceases to be a good measure. (Strathern, 1997)

Our target article set out with an ambitious goal: To provide a unified account of the disparate descriptions of proxy failure across not only social and natural sciences but also engineering and artificial intelligence research. Clearly such a task could only truly be achieved by a broad transdisciplinary community. The wide range of commentaries from a remarkable array of disciplines has not only confirmed our main proposition – that the disparate cases do reflect the same core phenomenon of "proxy failure" – but suggests the viability of a new interdisciplinary undertaking that we might call "proxynomics." Several comments highlighted key issues and extensions, others provided novel illuminating examples and, perhaps most importantly, quite a few proposed viable solutions or at least mitigation strategies.

R1. Introduction

When we began exploring the idea that the disparate array of phenomena referred to here as *proxy failure* might reflect a single core mechanism, we were a little intimidated. It seemed obvious that there are deep similarities between descriptions in widely varying fields, and indeed many previous authors have highlighted these similarities (e.g., McCoy & Haig, 2020; Smaldino & McElreath, 2016; Zhuang & Hadfield-Menell, 2021). It also seemed clear that many of the issues discussed were of current scientific interest (e.g., honest signaling) or indeed pressing social importance (e.g., scientific replicability, artificial intelligence [AI]-alignment, economic growth-driven climate breakdown; see Table 1 of the target article).

But, if proxy failure was really, at its core, a single phenomenon, would not someone else have elaborated on this? Indeed, several separate disciplines had developed entire literatures which seemed to specifically address the issue of proxy failure, for instance principal-agent theory in economics, signaling theory in ecology, or AI alignment research. Clearly, countless scholars have thought long and hard about the issue. So perhaps the similarities that were so frequently noted were superficial. Perhaps we were succumbing to an overgeneralization bias (Peters, 2022). Or perhaps, trying to explain everything would end up explaining nothing. It also did not escape our attention that many of the issues we identified as proxy failure (e.g., honest vs. runaway signaling, revealed vs. normative preferences, teleonomy vs. teleology) remain highly contentious within disciplines. Perhaps it was sound academic intuition which had led previous scholars to avoid an attempt at theoretical unification.

Nevertheless, it occurred to us that barriers to transdisciplinary communication could also have prevented a unified account. Scholars are rightly careful when trying to interpret the concepts of a neighboring discipline. Often, these are built on a body of tacit knowledge and assumptions that is not fully understood by an outsider. Indeed, it took our team, composed of a neuroscientist, an economist, an ecologist, and a neuro/meta-scientist, over three years of almost weekly meetings to build the confidence that we are talking about the same things. Yet the litmus test remained to be conducted. Our account clearly required broad scrutiny from experts in each of the disciplines we drew on. Furthermore, if we had described a general and recurring pattern, then there would almost certainly be numerous examples of proxy failure beyond what we had considered. We are

Table R1 (John et al.). Overview of comments by field

Comment area	Commenter
Ecology	Nonacs, Harris, Bartlett
Decision making	Ainslie, Sacco, Robertson et al., Burns
Economics	Moldoveanu, Kapeller et al., Sacco
Education	Samuelsson, Haig
AI/cybernetics	Kurth-Nelson et al., Robertson et al., Pernu
Science	Watve, Kapeller, Browning & Veit, Higgins et al., Sadri
Society	Szocik, Fernández-Dols, Ullman & Bridgers, Kurth-Nelson et al.

gratified to find that the wide range of commentaries provides both in-depth scrutiny and numerous additional examples. Overall, we have received 20 comments by 35 authors in more than six fields (Table R1). These have highlighted numerous interrelated trade-offs affecting proxy failure (Table R2) and in more than one instance introduced intriguing novel "proxynomic" concepts (Box R1). In the following, we discuss both criticisms raised and additions made, organized around four broad themes:

1. On goals
2. Additional examples
3. Drivers
4. Proposed solutions

R2. On goals: Nested hierarchies from nature to society?

Several commenters raised potential problems which revolve around the identification or attribution of goals or "teleonomy." For instance, some argued that particular examples do not show proxy failure when a different entity within the system is designated as the agent or the regulator; or that if we only understood the goal correctly, we would see that there is no proxy failure at all (e.g., **Nonacs**). Other commenters argued that we relied too much (**Pernu**) or too little (**Bartlett**) on "anthropomorphic" notions of goal-directed behavior. Further comments discussed the difficult issue of defining human goals (**Ainslie** and **Sacco**), and in particular collective social goals (**Fernández-Dols**). In our view, these comments serve to reiterate one of our key points: Proxy failure (or success) is *always, and inevitably*, defined with respect to a specific assignment of a regulator, goal, and agent. Disputes about specific examples of proxy failure often stem from the assumption that there is a "correct" level of analysis or a single "correct" goal. Instead, recognizing teleonomy across a multiplicity of biological and social levels can help us understand the overall behavior of a biological system or indeed a society.

R2.1. What is a "correct" goal?

Nonacs argues that the "true goal" of a peahen is to maximize offspring fitness, *which itself depends on the proxy*. This is a completely valid, and indeed important, perspective. In this Prum-ian view, beauty is itself the important goal (because it makes for a "sexy son," Fisher, 1930; Prum, 2017). However, many biologists view honest signaling as the main driver behind beautiful traits – that is, beautiful ornaments evolve to prominently signal male quality *above and beyond* simply the odds of (a son) landing a mate. So yes, proxy self-referentiality (in this case across generations) implies that a proxy becomes a goal in itself. Similar runaway processes routinely happen in social systems (Frank, 2011). But we feel it is nevertheless illuminating to ask if the proxy reflects some "underlying" value or simply drives a runaway self-referential process. The interesting comment by **Harris** dissects the potential determinants of which of the two might apply, and seems well worth further developing. The reason is that the distinction matters. The issue is closely related to the potential conflict between the welfare of a population and that of the individual, which Frank (2011) called "Darwin's Wedge." Consider for instance, a neighboring population of peafowl (say on a different island), which for some reason does not slip into the runaway process (say a predator catches peacocks with above average tails). All other things being equal, this second population would spend less energy on wasteful signaling, and thus have greater average fitness. In this way Nonacs' comment highlights the hierarchical nature of many nested systems of proxies and regulators. What is good at one level, enabling a proxy-trait

Table R2 (John et al.). Proxynomic tradeoffs highlighted by commenters

Tradeoff	Example	Comment
Legibility vs. fidelity	Legible test scores, which are proxies for psychological constructs, get published in lieu of better validated constructs (proxy pruning).	Higgins et al.
Objectivity vs. subjectivity	Students demand exact curricula and "objective" evaluation criteria, giving them more opportunity learn to the test (i.e., hack more legible proxies).	Haig
Integration vs. diversity	Informational integration in AI-systems (or modern society), that is, the use of a single global proxy, promotes failure generally and may lead to catastrophic failure (Fermi-Paradox).	Kurth-Nelson et al.
Dynamism vs. statism	Dynamic change and continuous reappraisal of an objective function (i.e., a proxy) can help to mitigate proxy failure.	Kurth-Nelson et al.
Concentration vs. pluralism	Preferential attachment in citation networks may lead to excessive academic concentration, undermining pluralism and disruptive science.	Kapeller et al.
Consistency vs. inconsistency	Organisms pursue allostasis concerning multiple potentially inconsistent goals. Market environments optimize for a single consistent variable (amount consumed) leading to, for example, super-addictive substances.	Sacco
Extrinsic vs. intrinsic motivation	Intrinsic motivation (e.g., in child play) can be harnessed to further (educational) goals with decreased risk of proxy failure (the proxy represents an extrinsic incentive).	Samuelsson, Ullman & Bridgers, Haig

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Box R1 (John et al.). New *proxynomic* concepts from the commentary

- **Proxy pruning (Higgins et al.**): When less legible proxies (or aspects of a proxy) are successively pruned away, leaving only highly legible and communicable proxies which may however be poor approximations of the goal. Example given: The pruning away of construct validity controls in psychological research.
- **Proxy complementarity** (**Watve**): When more than one set of agents have an interest in hacking a proxy. Example given: Journals and authors have an aligned incentive to inflate the journal impact factor (JIF) when the latter publish in the former, and even funding agencies may benefit from the illusion of impact.
- **Proxy exploitation** (**Watve**, **Kapeller et al.**): When an established proxy in one system is exploited by an external system. Example given: Private academic publishers exploit the JIF to attain record profit margins of 20–40%, with little apparent contribution to the original goal (furthering science). Disambiguation: This seems closely related to our concept of proxy appropriation, in which we have however posited a nested system in which "a higher level goal" appropriates a lower level proxy.
- **Predatory proxy** (**Watve**): When a "proxy devours the goal itself." This seems well aligned with the concept of "goal displacement" according to which the proxy eliminates the old goal and becomes a new goal in its own right (Merton, 1940; Wouters, 2020).

to sweep through a population, might plunge it toward extinction (Prum, 2017, p. 131, also note **Kurth-Nelson et al.**'s reference to the Fermi Paradox). A cancer cell might fulfill its goal of unmitigated growth, but at the ultimate cost of the host organism. Or as **Pernu** writes for the case of addiction: "from the perspective of the dopamine system there's a success." Regarding sexual selection, Prum (2017, p. 133) calls this *aesthetic decadence* – every individual club winged manakin might have won mates with its club wings but this came at the cost of a decrease in overall survivability of the species. Likewise, the individual marsupial who developed large strong forearms as a neonate might have beaten out all its competitors but the spread of this trait may have permanently restricted marsupial mammals to limited evolutionary niches compared to mammalian radiations on other continents (no marsupial whales or bats). So yes, one can always characterize a system in terms of a goal for which there is no proxy failure. And the notion of "the good of the population" as a goal is rightly contentious. But a broader perspective – asking for instance why some populations or species exist or proliferate and others collapse or go extinct, or why *aesthetic decadence* seems pervasive in both animals and humans – reveals the epistemic power of rejecting the notion of a correct goal. However, we certainly agree with Nonacs' perspective that beautiful signals need not signal anything other than an individual's beauty.

R2.2. Goals in nature: Invalid anthropomorphism or underestimated reality?

Two comments, those of **Bartlett** and **Pernu**, dig deeper into the underlying philosophical question of whether it is appropriate to engage in the ostensibly anthropomorphic activity of identifying goals outside of conscious human behavior. Previous discussions had made us painfully aware that this issue would lead to controversy, in particular as we write about goals of presumably non-conscious social and biological systems and rather liberally use the concept of an agent. We were gratified to find that opposing academic poles, between which we attempted to tiptoe, found articulation in the comments by Pernu and Bartlett. On the one hand, Pernu argues that we have not gone far enough in "naturalizing" the phenomenon of proxy failure, suggesting we should avoid all anthropomorphisms (even for human systems?). He argues that, for a fully general account, speaking of goals and agents would be "misleading." On the other hand, Bartlett thinks that our "selectionist bent" does not do sufficient justice to teleonomic (i.e., clearly goal-directed) forces in evolution, which have been extensively examined under the heading of the "extended evolutionary synthesis." In our view, both comments add valuable perspective and are compatible with our account.

First, we want to applaud **Pernu** for asking whether we have gone far enough in identifying a fully general model. He proposes to use the language and tools of adaptive control theory and explore the parallels between engineering research and biology, which seems to us an excellent program of research. To "view all organisms and evolutionary processes as hierarchies of control systems," is precisely what we have attempted to do – the notion of nested hierarchies is central to sections 4 and 5 of the target article. This helps us emphasize yet again that proxy failure for one level can look like a success for another level. Interestingly, Pernu suggests that the "correct" level to assess the goal is always the "encompassing" system (would this be the species for **Nonacs**' case?). While this claim intuitively holds for examples like that of addiction (the higher level being the conscious individual and the lower level the dopamine system), there are also many cases where lower-level goals seem to be more relevant. For instance, economic/corporate goals and proxies (the "encompassing" social level) can subvert individual goals (conscious individual level) by hacking our dopamine system (sub-individual level, also see Box 2, 4, and 8 of the target article as well as the comments by **Sacco** and **Robertson et al.**). In this case, a higher-level proxy (profit) should arguably be subordinated to the lower-level goals of individual humans, which may not include being manipulated into an addiction for profit. Note that in this case the higher-level proxy (profit) leads to social teleonomic behavior, that is, the market will behave *as if* to pursue the higher-level goal of say maximizing GDP, and that this emergent higher-level goal cannot be assumed to be identical to the individual-level goals of market participants (Box 4 of the target article). A final point worth making is that, while Pernu sets out to extinguish anthropomorphism from our account, the very notion of "adaptive control" is in our view just another restatement of the same anthropomorphic concepts. What is the target state of an adaptive controller other than a teleonomic goal, or indeed (in engineering) an intentional human goal? In our view, a truly general account must include human goal-oriented behavior at both individual and social levels. So we wonder if Pernu is suggesting we should also avoid anthropomorphizing humans?

This leads us to **Bartlett**, whom we want to thank for pointing the reader to the extensive literature on teleonomy and its more recent developments in the "extended evolutionary synthesis" (Laland et al., 2015). Our thinking is so closely aligned with this literature that we struggle to see how exactly Bartlett thinks our account conflicts with it. He claims that our "selectionist bent" led to problems, but this seems to be based on several misunderstandings. Granted, our brief treatment of evolutionary biology did not do justice to the diverse forces that influence the natural world (drift, epigenetics, controlled hypermutation, etc.);

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

we had little choice but to focus on the areas where proxy failure and its constraints were easiest to illustrate. But we do explicitly discuss several concepts that fall squarely into the extended evolutionary synthesis such as the "evolution of beauty" through runaway sexual selection, or "runaway niche construction," both of which Bartlett himself highlights as paradigmatically teleonomic (Bartlett, Bartlett, & Jonathan, 2017). Further, it can be argued that the proxy-treadmill is a key mechanism for developing novel elaborate traits (evolutionary decadence), thereby accounting for far more than simplistic selectionist mechanisms based on honest signaling. Another place where Bartlett feels our "selectionist bent" led us astray is our statement that selection at a higher level is a "hard constraint" on proxy failure at a lower level. Yet the statement is difficult to dispute. If a peacock's tail grows *too* large, the peacock will not live. If a corporation maximizing KPIs remains unprofitable for two long, it will go bankrupt. None of this contradicts what we believe Bartlett actually wants to emphasize, namely that phenomena such as the "evolution of beauty" or "directed cultural niche construction" (Braganza, 2022) – which operate within the bounds of this hard constraint – can play a pivotal role in evolution because they shape what is later available for (potentially changing) selection. Furthermore, in the case of markets, the rules of selection are almost entirely socially constructed, and in a sense thus purely intentional. Again, this does not conflict with the observation that a given system of market selection provides a "hard constraint" on which types of firms can exist, or the degree to which proxies within them can inflate. Instead, it highlights that the way we set up selection (the proxy) matters profoundly. It is the close inspection of the proxy – the determinants of selection – which can allow us to judge whether or not our individual and social goals are served by market competition.

R2.3. Nested hierarchies of goals within humans and societies?

In biology, disputes about the "correct goal" seem to mostly entail conflicting interpretations. In psychological and social areas, such disputes tend to spill over into conflicts about which goals we *should* pursue. The transition between biological and human systems is notoriously contentious, as for instance illustrated by **Bartlett**'s admonition that we should not "equate" the intentional "selection of action" with non-intentional "selection of traits." We in no way meant to "equate" the processes, but only to point out a "functional equivalence," which seems to explain why proxy failure can occur in both domains. Such functional equivalence is fully compatible with the fundamental differences highlighted by Bartlett (though we do wonder about intermediate cases, e.g., in animal decision making or the immune system; Noble & Noble, 2018). Above all, this perhaps highlights the contentious transition between teleonomy and teleology. While we of course have no answers to this challenging philosophical issue, we do believe both the target article and the commentary highlight that it is often essential to jointly consider natural/biological and human/intentional proxy systems.

Both **Ainslie** and **Sacco** touch upon the transition between biological "*as if*" goals and genuine human goals. Ainslie points out that proxy-rewards in human brains must necessarily occur over much shorter timescales than evolution. Indeed, we can quickly and adaptively learn new proxies to guide our every-day behavior. In many cases, a metric that starts as a proxy for a bigger goal can become a goal in itself. Examples could include the pursuit of money – once a tool to achieve survival and reproductive objectives, now pursued by many people as a goal in itself. But the underlying point, to us, seems more profound, and closely linked to the historical distinction between teleonomy and teleology, or what one might call the emergence of *genuine goal-oriented behavior in humans*. Ainslie argues that "The self-selecting potential of proxies in neuroscience sets them apart from the other kinds of proxy the authors describe." Proxies within the brain "may compete like fiat currencies," that is through a self-referential process that frees them "from the need for predicting external rewards." A proxy may thus "turn into a goal in its own right." In the words of the neuroscientist Mitchell (2023, p. 68), the "open-ended ability for individuals to learn, to create new goals further and further removed from the ultimate imperatives of survival […] and ultimately to inspect their own reasons and subject them to metacognitive scrutiny" seems to underlie "the kind of sophisticated cognition and agency" which "might qualify as free will in humans." In other words, the human tendency to create, represent, and *reflect* complex webs of proxies and goals may have furnished our ability to individually and socially pursue genuine goals that have become distinct from evolutionarily programmed dictates. This ability to define and deliberate our own goals (and how to achieve them) is the foundation of the classical liberal conception of the rational agent which underlies not only economics but also democracy. In a sense, the free, rational agent of classical liberalism isolates this reflective ability and declares its biological underpinnings to be beyond inquiry – a given fixed utility function which we need only to optimize without questioning its origin or nature.

This is the starting point for **Sacco** who highlights that "a key source of proxy failures in economics stems from a mismatch between the evolutionary function of biological reward systems and traditional conceptions of utility maximization." He continues to note that "the unprecedented production and marketing of super-addictive goods exploits vulnerabilities in dopaminergic reinforcement learning circuits." **Robertson et al.** provide a similar view in the context of social media companies, which hack our attention allocation systems by presenting "threatening stimuli." This aligns well with the notion of our target article, whereby market economies, which are automated optimization devices, can undermine individual goals by hacking our decision-making systems. Sacco suggests that we can overcome such problems by redesigning our economic theory around concepts from biology (allostasis), neuroscience (reward prediction errors), machine learning (multicriterion optimization), and psychology (motivational heterogeneity). We enthusiastically applaud this vision, though it (much like our article) provides little detail on how exactly proxy failures through the economic provision of super-addictive products might be mitigated.

Another of **Ainslie**'s formulations naturally links individual goals to social goals: "A proxy that is a *good story* may turn into a goal in its own right (emphasis added)." "Good stories" can become "cultural attractors" (Falandays & Smaldino, 2022; Jones & Hilde-Jones, 2023) – shared narratives that can establish a cohesive societal goal across countless individual minds. This understanding seems closely related to that of **Fernández-Dols**', who introduces a dark twist. The commenter suggests that a shared narrative's purpose is not a seeming social goal like "justice," but only the "illusion of accomplishment" toward this goal, which allows the proxy (the narrative that something is more or less just) to self-perpetuate. The proxies become self-fulfilling and self-sustaining, quite independent of the abstract goals they purport to further. This is a profound insight which we have already highlighted in

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Braganza (2022), and indeed prominent sociologists have long before us (e.g., Luhmann, 1995). In social systems, proxies very frequently take on the role of perpetuating power structures. The approximation of some abstract, unattainable goal is only required as an illusion to "legitimate" the proxy, and thereby the power structure, to the masses. But, **Fernández-Dols**' conclusion from this observation, namely that the goals rather than the proxies are the problem, seems to go decidedly too far. Yes, the social proxies for abstract goals like "liberty, justice, wealth, hygiene, safety" will always remain imperfect and the goals themselves unattainable. We would also not dispute the author's claim that the pursuit of unattainable goals can be disheartening or even damaging. But it is key to recognize that *if we do deem a goal worthy*, then a proxy can be both a means of social legitimization *and* an approximation of an unattainable ideal. In this reading, the task is precisely to identify where proxies exclusively provide "illusions of success" or "social legitimations" and work to make those successes more real and the proxies more legitimate. The overwhelming majority of commenters seem to agree, since they suggest ways to further the goals, rather than abandoning them (e.g., **Samuelsson**, **Watve**, **Haig**, **Szocik**, **Browning**, **Kapeller et al.**, **Higgins et al.**, **Sadri et al.**, **Kurth-Nelson et al.**, etc.). That said, we must of course recognize that many goals, Fernández-Dols names "recreating France in colonial Vietnam," are ill-conceived and worth abandoning. Examining whether a goal is really worth pursuing is of course essential, and we must thank the author for emphasizing this.

R3. Additional examples: Open and closed proxy failure and non-scalar proxies

Numerous commenters suggested additional examples of proxy failure, strengthening our intuition that the phenomenon manifests far beyond what we were able to review. In order to contextualize these additional examples, we feel the need to introduce a distinction between what we will term *open* and *closed* proxy failure. Open proxy failure simply describes the existence of a poor proxy – the proxy fails in the sense that it does not allow the achievement of the goal for reasons that are external to the proxy-based optimization process. By contrast, *closed* proxy failure involves a proxy that becomes worse *because of* proxy-based optimization. Our initial definition was restricted to closed proxy failure – we will borrow the words of **Pernu** who concisely recapitulates this: "proxy failure [is not] a mere signal failure: central to the analysis is that the target system hacks the proxy." However, as **Kapeller et al.** rightly state, this is "not the only way in which proxies can fail." Some examples from **Szocik**, **Watve**, **Kapeller et al.,** and **Higgins et al.** appear to reflect instances of open proxy failure, that is, cases that go beyond our original definition. Given that so many authors intuitively apply the term to both open and closed cases, it seems more appropriate to introduce the distinction than to insist on our original definition. Nevertheless, the distinction is key because it can lead to diametrically opposing implications: In open proxy failure, a greater reliance on the proxy may help to achieve the goal, while for closed proxy failure, it will tend to make the problems worse.

Finally, **Burns**, **Ullman & Bridgers**, and **Nonacs** suggest extensions or additional examples, of which we remain unsure if they fit well within the present framework. This is not to say that they are not related to it or interesting in their own right. But expanding the scope too wide comes at the cost of decreasing conceptual precision and clarity.

R3.1. Open proxy failures

Both **Kapeller et al.** and **Watve** explicitly highlight the distinction between open and closed proxy failure. They refer to "the political economy of scientific publishing" (Kapeller et al.) as a key illustrative example of what Watve calls "proxy exploitation." Specifically, they note that the proxy of one system (journal profitability in the economic system) is simply not geared toward the goal of another system (advancement of knowledge in academia). In the words of Watve: "Editorial boards appear to strive more for journal prestige than the soundness and transparency of science (Abbot, 2023). More prestigious journals often have higher author charges (Triggle, MacDonald, Triggle, & Grierson, 2022) and make larger profits with little contribution to the original goals." This is best construed as an instance of open proxy failure because it is not the pursuit of the proxy within the academic system that leads to failure. Indeed, the notion that a well-established proxy in one system could be exploited by another system or party with distinct goals seems well worth the additional label of "proxy exploitation" supplied by Watve. For instance, the phenomenon of mimicry among prey species, supplied by **Nonacs**, seems to fit the label proxy exploitation well. In it, a dangerous or poisonous prey species advertises this danger to a predator through bright colors or high visual contrasts (a proxy signal). This allows a second prey species (which is not dangerous) to mimic the proxy, and exploit the proxy-system, in order to repel predators. In contrast to our concept of proxy appropriation, in which a lower-level proxy is appropriated by a higher-level goal, proxy exploitation could refer to any case in which the systems are not nested.

Another example provided by **Kapeller et al.** is the academic "third mission," which can be summarized as unfolding societal impact. Typical proxies for this, such as the number of public appearances or so-called alt-metrics, may simply fail to capture broader societal goals. This may be because the proxies are poor, or it may reflect an instance of "proxy exploitation," in which the proxies of one system (such as social media engagement) fail to reflect the goals of another system (dissemination of accurate knowledge by academia).

The tension between open and closed proxy failure is also apparent in **Szocik**, who criticizes the use of *quotas* as proxies for "antiracism" and "antisexism." We should begin by applauding Szocik for drawing attention to this important topic. Szocik explains why quotas are not *sufficient* to eliminate all the subtle drivers and determinants of racism or sexism. He highlights, for example, that quotas in some western power structures will do nothing to alleviate sexism against millions of women around the world. Furthermore, limited quotas tend to help a small non-representative subgroup of a minority rather than those most discriminated against. For instance, antiracist quotas in US universities tend to lead to the recruitment of highly educated first-generation immigrants rather than the descendants of former slaves. However, we feel it is important to highlight that these arguments imply only "open proxy failure" and thus do not help to predict if more or stricter quotas would improve the situation. For instance, the observation that proxies do not help where they are not applied seems trivial, and clearly suggests wider application rather than abandonment as a solution.

That said, we would acknowledge several ways in which **Szocik** suggests quotas could indeed also drive closed proxy failure. First, quotas could create sexist or racist backlashes, for instance when individuals are dismissed as having gained their positions not via merit but via quotas. Second, quotas can create an "illusion of success", as highlighted by both **Fernández-Dols** and Szocik. Both aspects are clearly worth sustained attention. Nevertheless, it seems to us that the overall evidence strongly suggests quotas do help (Bratton & Ray, 2002; Chattopadhyay & Duflo, 2004). Representation does not, alone, solve racism and sexism, but it is one important component of a solution.

Another example of open proxy failure (in our reading) is provided by **Higgins et al.** The commenters note that "constructs" in psychological research are typically assessed via forms of standardized tests, which are then used to create a score which stands as proxy for the presumed psychological trait (consider for instance the controversial IQ score; Gould, 1996). Whether or not such a score is a good approximation of an underlying construct is then expressed under the heading of "construct validity," which psychologists have developed sophisticated methods to assess. However, Higgins et al. note that "there is a growing body of evidence demonstrating that studies across psychological science routinely accept measurements as valid without sufficient validity evidence […], including measurements used for important clinical applications." In other words, psychologists routinely use poor proxies, even though they should know better. The reason this could be classified as open proxy failure is because the limitations of the proxy seem to derive primarily from limits to the legibility of the underlying construct, and it is not the psychologist's pursuit of an accurate understanding of this construct that leads to failure.

R3.2. Closed proxy failure

Notably, a subtle shift of frame recasts the example described by **Higgins et al.** as closed proxy failure. Specifically, if the goal is defined as *the scientific communication of valid constructs*, then this may cause a "prioritisation of legibility over fidelity." In this framing, it is ultimately academic competition for publication space (which hinges on highly legible test-scores) that undermines the validity of those scores in capturing the purported underlying constructs. Higgins et al. introduce the intriguing concept of "proxy pruning" to describe how this might occur in practice: Difficult qualitative questions about construct validity may be successively pruned away, to the benefit of simple legible metrics (such as test scores). Unproven metrics are often considered more valid if they correlate with other, also unproven but older, metrics. Similarly, **Sadri & Paknezhad** show that drug discovery operates by researchers trying to find molecules that bind to a target protein, rather than drugs which affect the human health phenotype. Huge biotech companies are built on this "target-based drug discovery" method, which is a reductionist method that is highly legible but of questionable validity. **Browning et al.** describe how animal welfare is typically measured through biochemical proxies of stress, such as cortisol – which is in fact not even an accurate metric of emotional state (stress can be negative fear, positive arousal, etc.). Instead, they argue for more complicated, but accurate, measurements of actual emotional state. We have already mentioned **Szocik**, who argues that many highly legible antisexist and antiracist policies fail to achieve the actual goal, equality. In each of these cases, proxy pruning may translate good legibility to proxy failure.

Several other comments presented examples of closed proxy failure in academia. Indeed, **Watve** admonishes that we did not "adequately cover proxy failure in academia" where "proxy failure has reached unprecedented and unparalleled levels." This is perhaps the right place to disclose that proxy failure in academia was in fact one of the main areas of research that lead up to the current target article (Braganza, 2020, 2022; Peters et al., 2022). So our main justifications for neglecting the topic in the present article are that (i) we have written about it previously and (ii) we felt sure it would be raised by others. Indeed, not only Watve, but also **Kapeller et al.**, **Higgins et al.**, **Sadri & Paknezhad**, and **Browning et al.** addressed proxy failures in academia. Watve's point that studying academia has particular potential to propel proxynomics forward is well taken. In addition to the above-mentioned concept of *proxy exploitation*, he proposes the concept of *proxy complementarity*, which again seems highly generalizable. *Proxy complementarity* suggests that proxy failures can accelerate (or become entrenched) when several parties stand to mutually gain from hacking (e.g., researchers and journals have a joint incentive to inflate Impact Factors).

We have already mentioned two examples of open proxy failure introduced by **Kapeller et al.** In addition, they describe an example of closed proxy failure in academia. Specifically, they highlight that large and established fields (or institutions) are more attractive for new entrants, or more impressive for evaluation agencies. This can lead to a preferential attachment dynamic – the proxies of size and prestige lead to a lack of diverse and disruptive science, which many evaluation agencies or researchers may claim as their actual goal.

In summary, mechanisms like *proxy pruning* or *proxy complementarity* likely foster closed proxy failure in many domains. Proxy-usage favors more legible over more accurate proxies. The intensive use of the proxy then subsequently causes failure along all the dimensions in which fidelity was sacrificed for legibility. Having more (rich white) women CEOs, (sad) cows with low cortisol, (useless) papers on candidate drug therapies and diabetes, and (invalid) metrics of psychological capabilities may be counterproductive to the real goals of equality, animal welfare, successful medical therapies, and rich understandings of our minds because we feel complacent, having checked the box. More sinister cases, as outlined for example by **Watve**, involve direct intentional gaming and are thus also clear cases of closed proxy failure.

R3.3. Unclear cases

Finally, several authors have suggested extensions of our framework about which we remain unsure.

Nonacs proposes the lack of a reliable paternity signal among animals as an instance of proxy failure. Such a proxy would be a "greenbeard" signal (Haig, 2013) that could prove advantageous to both father and offspring by allowing fathers to allocate parental efforts only to their own offspring. Nonacs thus raises the question of why it is not common, suggesting it could be a risk-tradeoff between being nurtured by the own father versus being killed by another's father. In our view, since the lack of a genetic proxy here would not be due to any failure, but due to *the advantage of not signaling*, this example does not qualify as proxy failure.

Burns as well as **Ullman & Bridgers** suggest we could extend the notion of proxy failure to non-scalar proxies, namely heuristics and utterances, respectively. Although both contributions are highly interesting in their own right, we feel that this expansion in scope might come at too great a cost. Defining proxies as scalars allows a researcher to map proxy failure to mathematical approaches to optimization within several disciplines – something we view as highly desirable. Moreover, the scalar nature of proxies

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

appears to be central to the amplifying effect of regulatory feedback – a point we discuss further below. But this in no way diminishes the insightfulness of the comments, which do highlight clear similarities.

R4. Additional drivers of proxy failure

Several commenters outlined additional drivers of proxy failure, which are likely to be of some generality.

R4.1. Sophisticated agents

Moldoveanu proposes that proxy failure occurs because agents are generally more sophisticated than regulators concerning the relevant task. This is likely true in many cases, and for these the "explanation-generating engine" outlined by Moldoveanu should be very valuable. In particular, the notion that agents typically will (i) see more, (ii) spend more time, and (iii) think more, concerning the particular task at hand, and thus exceed the regulatory model's sophistication for it, seems plausible. However, we feel it is worth reemphasizing that agents do not need to be sophisticated *at all* for proxy failure to occur. This is illustrated by the examples in our target article in which agents are passive recipients of selection, such as in many ecological cases, in the neuroscientific case, and in machine learning. Even in social and economic cases it may often be more accurate to think of proxy failure as a passive selection phenomenon, rather than a consequence of devious hacking by agents (Braganza, 2022; Smaldino & McElreath, 2016). We would also question Moldoveanu's concluding notion that "the desire of agents to 'one day' themselves become principals and regulators" can help mitigate proxy failure. He seems to follow a similar notion as **Nonacs**, who argues that the fact that marsupials who win the race to the teat will some day later themselves become mothers, would somehow constrain proxy failure. Instead, we would argue that the rational agent (or the rational marsupial) must concentrate on maximally gaming the system while young in order to become a regulator. This may reduce overall fitness and may well sometimes lead to catastrophic failure, but if higher-level selection is sufficiently slack, both firms and species will persist despite wasteful proxy-games. As Frank (2011) has argued, all marsupials would be better off if they could form a contract to halve their front paw size – because proxy value is relative to that of competitors, this would leave the outcome of competition completely unaffected, while saving everyone lots of energy. Of course, marsupials (or in Frank's original example, deer) are not known for their contracting skills. But more sophisticated agents can be better at this, as in the case of athletes' organizations that regulate the terms of their competition by banning doping. Notwithstanding such observations, which indicate that agent sophistication can both amplify and reduce proxy failure depending on specific factors, Moldoveanu's general insight, that there may be a sophistication gradient between agent and regulator, strikes us as worth exploring further, particularly in economic contexts.

R4.2. Proxy legibility

Several commenters have outlined how proxy legibility can drive proxy failure. This is an intriguing and perhaps ironic insight in a world that increasingly relies on objective, quantitative, and legible measures of *accountability*. In this context, it may be worth highlighting that the most famous and pithy formulation of the phenomenon of proxy failure – "when a measure becomes a target, it ceases to be a good measure" – received its name in an essay by Hoskin (1996) entitled "The awful idea of accountability," and received its final phrasing by Strathern (1997) in a critique of the "audit explosion" in higher education.

We have already introduced **Higgins et al.**'s proposed mechanism of proxy pruning. It could happen passively, simply because more legible proxies are easier to communicate. But legibility could also distract the attention of regulators from the task of assessing the agreement between goal and proxy. Similarly, **Haig** has highlighted in the context of education that more legible, or more "objective" evaluation criteria are also far easier for students to hack than "subjective" evaluations. He notes that "there may be more striving for excellence when an agent does not know how their work will be judged." And indeed, both "revealed preference in economics, and fitness in evolutionary biology" have "low legibility before judgment is pronounced." Resilience toward proxy failure may therefore derive from limited legibility.

R4.3. Proxy complementarity

Watve highlights a key property that can drive proxy failure, which he calls "proxy complementarity." It applies when more than one set of agents have an interest in hacking a proxy. A prominent example, one that incidentally was also highlighted by **Kapeller et al.**, is the following: Journals and authors have an aligned incentive to inflate a journal's impact factor when the latter publish in the former. Indeed, even funding agencies may benefit from the illusion of impact. It appears like most of the relevant players have an overt proxy incentive to allow inflation (or to aid other parties in hacking the proxy). This could be argued to have led to a situation that Merton (1940) called "goal displacement" and Watve calls "predatory proxy": The proxy has become so dominant as an evaluation tool that it has begun to replace the underlying goal. Nevertheless, the very fact that countless voices are discussing the issue arguably proves that goal displacement has not been completed in academia. Regardless, the key point here is that proxy complementarity may be a powerful driver of proxy failure.

R5. Proposed solutions

Perhaps the most natural question to ask when confronted with the issue of proxy failure is: "How do we solve or at least mitigate it?" This question already featured prominently in our article, where, for instance, we suggested preemptive action by managers, or higher level selection, as mechanisms that can mitigate proxy failure. However, we were thrilled by the depth and quality of additional mitigation proposals offered by, among others, **Kurth-Nelson et al.**, **Samuelsson**, **Burns**, **Sacco**, and **Ullman & Bridgers.**

R5.1. Emphasizing intrinsic incentives and play

In human systems, a group of mitigation strategies for proxy failure revolves around the intrinsic incentives of the agents.

Samuelsson highlights the under-studied role of "play" in human and animal education, offering it as a paradigmatic solution to avoid proxy failure: Play is an intrinsically joyful activity, which greatly enhances education, seemingly as a by-product. Proxy failure explains why, "paradoxically, relying solely on educational goals can lead to worse learning." Students seem to succeed best when they are driven by free and intrinsic motivation. The same notion is captured by **Haig** when he writes, "Monet and Picasso did not paint by numbers." Such a view highlights how counterproductive it might be for schools in the USA to cut recess in order to free up more time to teach to the test! **Samuelsson**'s assessment, that play is an antidote to proxy-failure-laden socio-educational-economic pursuits, seems wholly plausible to us. He describes play as "loosening" educational proxies and argues for a world where we embrace looseness to avoid proxy failure. A further key insight is that "as children play with and socialize around the tools and technologies in their environment (Samuelsson, Price, & Jewitt, 2022)," play allows highly flexible or "culturally appropriate learning" even in rapidly changing cultural environments. This hints at another way in which play might award resilience toward proxy failure. That play has no measurable proxy for educational success is, in fact, a strength; we suggest that practitioners look for analogs of play elsewhere. For example, we have noticed among ourselves and our students that side projects – those pursued for joy rather than for credit, grants, or a degree – can be richer, more interesting, and more quickly completed than "main" projects. Of course, abandoning educational proxies altogether will likely prove counterproductive in a world where children need to meet basic standards concerning literacy or numeracy. Here, both sequential and parallel approaches to proxy diversity, discussed in more detail below, may come in handy. A sequential approach might involve the regulator interleaving proxy-oriented epochs with time dedicated to activities that do not count toward any proxy. A parallel approach might involve dividing up classrooms into teams based on specific projects, student preferences, and/or randomly assigned proxies. A regulator employing such approaches would need to be a playful "fox" rather than a focused "hedgehog" when it comes to the use – or avoidance – of proxies.

Several commenters more directly suggested that proxy failure can be avoided by aligning regulation with intrinsic agent goals and motivations. **Ullman & Bridgers**, in their charmingly written comment, suggest that "highlighting common-ground" can mitigate proxy failure where it is wholly or primarily caused by an agent *intentionally misunderstanding* the regulator's goal. For example, the rat-breeders in Hanoi knew what they were doing and so too does a "smart-ass" child following the letter, but not the sprit, of a parental diktat. **Fernández-Dols** seems to push toward a similar conclusion when he highlights an experiment in which gaming "was a consequence of the participants' perception of the [regulator's] instructions as an arbitrary imposition" and disappeared when "instructions were clear and procedurally fair." **Burns** argues that medical heuristics may be superior to proxies largely because they are often used in cases where the goal of the regulator and agent are aligned. Heuristics then simply turn hard questions into easy ones: "is this patient having a heart attack" becomes a list of three simple questions for doctors to triage patients. Hospital regulators want to regulate their doctors, but doctors already want to accurately diagnose heart attacks. We fully agree with these notions: Where common-ground can be found, it should be emphasized. Fostering the "intrinsic motivation" of agents – for instance by encouraging "peer norms and professionalism" (Muller, 2018, p. 112) – is preferable to extrinsic proxy incentives for another reason. It allows the agent to actually be a free agent, rather than an entity to be controlled. As **Haig** states "a subject's power to make decisions that others must accept on trust is a measure of the subject's sovereign agency." Social regulators rarely perceive the insult that is implied, when they impinge on this "sovereign agency." The insult lies not only in not trusting professional decisions, but in the tacit assertion that for example, doctors don't intrinsically want to help their patients or scientists are not intrinsically motivated to do good science (Muller, 2018). The upshot is that using proxies can do harm beyond the harm to the goal, which we referred to as (closed) proxy failure, namely it can demotivate the agents.

A more subtle instance of relying on intrinsic incentives is provided by **Robertson et al.**: Social media users' intrinsic incentives, almost by definition, reflect their goals. However, social media companies appear to have hacked user's decision-making apparatus to undermine these incentives by showing threatening stimuli. Humans instinctively attend to such stimuli, such that company proxies of user goals, namely watch time or click-rate, can be increased by presenting threatening stimuli. The case mirrors that of addictive substances – users engage in behaviors which they claim not to want and which demonstrably harm their wellbeing. To better align user goals with behavior, Robertson et al. propose that companies allow users to explicitly state their respective preferences about what they would like to see. For example, companies could implement an easy to access button labeled "I don't want to see content like this." Of course, companies may not do this because it will not make them money (here we have shifted frames, arguing that the corporate goal is not user-satisfaction but profit). In the terminology of our target article "a higher-level proxy of profit may constrain their ability or desire to further their customers goals." But this in no way invalidates Robertson et al.'s proposal as a step to solving the proxy failure they highlight: Take steps to align the regulator's goal with the agent's goal; give consumers the chance to opt out of certain content types; let email users actually unsubscribe from spam easily and permanently. If such actions do not currently align with a company's (profitability) goals, then an altered regulatory framework or better consumer education may be necessary.

A final point worth making in this context is of course that there may often be no common ground to seek, or no intrinsic incentive that aligns with the regulator's goals. In many ways this is the standard assumption of most economists, as perfectly illustrated by the comment of **Moldoveanu** (though there are of course exceptions; Baker, 1992; Holmstrom & Milgrom, 1991). Here, the notion that agents could in any way care about the regulator's goal simply does not feature in analysis. While we certainly feel **Ullman & Bridgers**, **Fernández-Dols**, **Burns** or **Kapeller et al.** are correct when they highlight the importance of intrinsic agent incentives (particularly in areas like parenting, medicine, or science), Moldoveanu's approach is clearly also appropriate in many situations. Some jobs will not be done because of intrinsic motivation, and here we will have to deal with extrinsic proxy incentives, and the ensuing failures. It is perhaps also worth reemphasizing that this best reflects most nonhuman instances of proxy failure, in which it often makes no sense to speak of intrinsic incentives toward the regulator's goal.

R5.2. Multiple proxies and "dynamic diversity"

Another fundamental mitigation strategy for proxy failure, outlined by **Kurth-Nelson et al.** and **Sacco**, seems to clearly apply in natural systems (Ågren, Haig, & McCoy, 2022), is increasingly explored in social systems (Bryar & Carr, 2021; Coscieme et al., 2020), and is now being formalized at a fundamental level in AI research. It entails the use of an array of heterogeneous proxies, together with dynamic evaluation schedules and continuous reappraisal – Kurth-Nelson et al. use the term "dynamic diversity." Some strands of reinforcement learning research suggest that multi-objective approaches can mitigate overfitting to a single objective (Dulberg, Dubey, Berwian, & Cohen, 2023; Hayes et al., 2022). Kurth-Nelson et al. also point to the "rich tradition" studying how diversity can mitigate proxy failure in the social sciences literature (Ostrom & Walker, 1994). All this clearly resonates with the comments by **Samuelsson** and Sacco who highlight motivational diversity and the *absence* of a single scalar to optimize.

Before we proceed, we must address a key issue about the nature of proxies. In the target article, we claimed that a proxy which a regulator uses to rank competing agents can typically be described as a one-dimensional quantity, that is, a single scalar. This requirement is inherent to *consistent* ranking: Scalar values map to unambiguous orderings. Such ordering and scalarity is implicit in abstractions such as *fitness*, on which natural selection is based, as well as *utility*, on which most of economics is built. Similarly, in reinforcement learning, there is a reward which is apportioned in a scalar fashion, necessitating a scalar proxy to guide credit-assignment. We have argued that in the brain, a course of action that produces higher amounts of a scalar proxy (such as dopamine bursts) receives higher amounts of regulator feedback (such as synaptic weight enhancements). In this view, regulatory feedback, whether it takes the form of rank-based selection or reward-like feedback, depends on the use of a single scalar proxy. The fundamental underlying claim is that, *to the degree that a system produces consistent decisions, even a multi-proxy optimization can be redescribed as a function producing one final scalar*. The upshot is that, even if our final proxy is a complex aggregation of many input proxies and factors (take, e.g., the Journal Impact Factor, where this is undeniably the case), this does not eliminate the risk of proxy failure. We stand by this claim.

But **Kurth-Nelson et al.** and **Sacco**'s point nevertheless stands. Firstly, even if diversity (say of the type we see in multi-objective optimization) does not fully "solve" proxy failure, it can clearly mitigate the risk or slow its time course. Integrating more proxies, even if they are formally equivalent to a single meta-proxy, increases the probability that weaknesses of individual inputs are counterbalanced. Secondly and more fundamentally, Kurth-Nelson et al. and Sacco reject a key premise on which we based our claim of scalarity – they *drop the requirement of consistency*. Distinct proxies, which are simultaneously used to guide behavior, may act at cross-purposes to each other and Sacco emphasizes that this closely resembles biological allostatic principles. Indeed, in an underappreciated study, Kapeller et al. (2013), building on Arrow (1950), derive the "impossibility of rational consumer choice" from the premise that consumers use multiple incommensurable proxies. "Impossibility" in this context means that, if incommensurable proxies lead to distinct rankings, then they *cannot be integrated into a single scalar*. For instance, if humans have distinct proxy systems for optimizing uptake of fluids and carbohydrates, and these proxies lead to distinct rankings of products, then no single scalar utility function exists over all products. Cutting edge research from reinforcement learning (e.g., Dulberg et al., 2023) shows that using multiple parallel proxies, which are not required to be consistent, can indeed improve performance particularly in changing environments. Another advantage was that it led to a natural exploratory tendency, reminiscent of biological organisms. Further, the alteration of proxies on a timescale much slower than that of individual regulatory decisions might create epochs that are long enough for agents to boost the production of each proxy, but not hack it. The system remains in the sweet-spot below over-optimization or "overfitting." Kurth-Nelson et al. bring up another interesting possibility related to proxy diversity: A subagent- or team-based "division of proxy labor." Agents could be divided into teams or sub-groups each incentivized to pursue a distinct proxy in parallel. In both the sequential and parallel approaches to proxy diversity, regulatory feedback must be more active. Some isolation of each team may be required to prevent unwanted combining of proxies. These and other forms of diversity may allow systems to avoid recapitulating proxy failure in new "meta-proxies." The usefulness of isolated subagents or teams raises additional questions that warrant further study: How does information-sharing influence how agents pursue proxies? Natural systems are never perfectly modular, which implies that information cannot be fully hidden from agents. How might this affect how agents balance the pursuit of proxies with the pursuit of their own goals, which may be distinct from those of the regulators? Even with diverse teams or partially inconsistent proxies, we suspect that no globally optimal strategy will be possible: Context-specific tradeoffs will inevitably be required between interpersonal legibility and motivational heterogeneity. These considerations point to the subtlety and complexity that successful mitigation strategies will likely require.

On a more general note, it seems almost certain that a wealth of research from across the disciplines touched upon will already have examined questions like those raised above. But integrating this knowledge remains a formidable task, which seems to require an entirely new level of trans-disciplinary scholarship. The remarkable reccurrence of proxy failure across all manner of natural and social systems does raise the question whether it is something akin to a natural law of goal-oriented systems, which leaves its traces even if it does not come to realize. It also raises deep philosophical questions about the nature of goals and agency and whether the traditional academic segregation between the humanities (including economics) and the natural sciences (in particular biology) can be sustained.

Rather ominously, **Kurth-Nelson et al.** close by suggesting that global catastrophic proxy failure (due to an excessive integration of "information societies") may be the reason for the Fermi Paradox. Fermi wondered why, given the infinitude of planets, we have not yet been contacted by extraterrestrial life forms. The commenters seem to suggest that advanced civilization may tend to self-annihilate due to planetary proxy-failure. Kurth-Nelson et al.'s speculation reminds of other dire warnings, such as Bostrom's (2014) famous paperclip factory, or indeed less theoretically, the notion of GDP-growth-driven planetary ecological collapse (Diamond, 2004; Kemp et al., 2022; Pueyo, 2018). We cannot comment further on such issues here, but we do think they serve to emphasize the potential scope and relevance of a trans-disciplinary study of proxynomics.

References

Abbott, A. (2023). Strife at eLife: Inside a journal's quest to upend sciencepublishing. *Nature*, *615*(7954), 780–781. https://doi.org/10.1038/d41586-023-00831-6

Ågren, J. A., Haig, D., & McCoy, D. E. (2022). Meiosis solved the problem of gerrymandering. *Journal of Genetics*, *101*(2), 38. https://doi.org/10.1007/s12041-022-01383-w

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, *58*(4), 328–346. https://doi.org/10.1086/256963

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, *100*(3), 598–614. https://doi.org/10.1086/261831

Bartlett, J., Bartlett, & Jonathan. (2017). Evolutionary teleonomy as a unifying principle for the extended evolutionary synthesis. *BIO-Complexity*, *2017*(2), 1–7. https://doi.org/10.5048/BIO-C.2017.2

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

https://doi.org/10.1017/S0140525X23002753 Published online by Cambridge University Press

Braganza, O. (2020). A simple model suggesting economically rational sample-size choice drives irreproducibility. *PLoS ONE*, *15*(3), e0229615. https://doi.org/10.1371/journal.pone.0229615

Braganza, O. (2022). Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, *9*(2), 1–41. https://doi.org/10.1098/RSOS.211030

Bratton, K. A., & Ray, L. P. (2002). Descriptive representation, policy outcomes, and municipal day-care coverage in Norway. *American Journal of Political Science*, *46* (2), 428–437. https://doi.org/10.2307/3088386

Bryar, C., & Carr, B. (2021). *Working Backwards—Insights, Stories and Secrets from inside Amazon*. Macmillan.

Chattopadhyay, R., & Duflo, E. (2004). Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica*, *72*(5), 1409–1443. https://doi.org/10.1111/j.1468-0262.2004.00539.x

Coscieme, L., Mortensen, L. F., Anderson, S., Ward, J., Donohue, I., & Sutton, P. C. (2020). Going beyond gross domestic product as an indicator to bring coherence to the sustainable development goals. *Journal of Cleaner Production*, *248*, 119232. https://doi.org/10.1016/J.JCLEPRO.2019.119232

Diamond, J. M. (2004). *Collapse: How societies choose to fail or succeed*. Penguin Books.

Dulberg, Z., Dubey, R., Berwian, I. M., & Cohen, J. D. (2023). Having multiple selves helps learning agents explore and adapt in complex changing worlds. *Proceedings of the National Academy of Sciences*, *120*(28), e2221180120. https://doi.org/10.1073/pnas.2221180120

Falandays, J. B., & Smaldino, P. E. (2022). The emergence of cultural attractors: How dynamic populations of learners achieve collective cognitive alignment. *Cognitive Science*, *46*(8), e13183. https://doi.org/10.1111/cogs.13183

Fisher, R. (1930). *The genetical theory of natural selection*. Clarendon Press. https://books.google.com/books?hl=de&lr=&id=WPfvAgAAQBAJ&oi=fnd&pg=PA1&ots=onzSFD_Yvp&sig=r7RthPTyPdIjR2p4PCpeIHvxnNg

Frank, R. H. (2011). *The Darwin economy: Liberty, competition, and the common good*. Princeton University Press.

Gould, S. Jay. (1996). *The mismeasure of man*. Norton. https://www.amazon.de/Mismeasure-Man-Stephen-Jay-Gould/dp/0393314251

Haig, D. (2013). Imprinted green beards: A little less than kin and more than kind. *Biology Letters*, *9*(6), 20130199. https://doi.org/10.1098/rsbl.2013.0199

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., … Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, *36*(1), 26. https://doi.org/10.1007/s10458-022-09552-y

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7* (January 1991), 24–52.

Hoskin, K. (1996). The "awful idea of accountability": Inscribing people into the measurement of objects. In R. Munroe & J. Mouritsen (Eds.), *Accountability: Power, ethos and the technologies of mana* (pp. 265–282). International Thomson Business Press.

Jones, J. H., & Hilde-Jones, C. (2023). Narrative as cultural attractor. *Behavioral & Brain Sciences*, *46*, 1–74. https://doi.org/10.1017/S0140525X22002667

Kapeller, J., Schütz, B., Steinerberger, S., Kapeller, J., Schütz, B., & Steinerberger, S. (2013). The impossibility of rational consumer choice A problem and its solution. *Journal of Evolutionary Economics*, *23*, 39–60. https://doi.org/10.1007/s00191-012-0268-2

Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., … Lenton, T. M. (2022). Climate Endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, *119*(34), e2108146119. https://doi.org/10.1073/PNAS.2108146119

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., … Odling-Smee, J. (2015). The extended evolutionary synthesis: Its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences*, *282* (1813), 1–14. https://doi.org/10.1098/RSPB.2015.1019

Luhmann, N. (1995). *Social systems*. Stanford University Press.

Mccoy, D. E., & Haig, D. (2020). Embryo selection and mate choice: Can 'Honest Signals' be trusted?. *Trends in Ecology and Evolution*, *35*(4), 308–318. https://doi.org/10.1016/J.TREE.2019.12.002

Merton, R. K. (1940). Bureaucratic structure and personality. *Social Forces*, *18*(4), 560–568. https://doi.org/10.2307/2570634

Mitchell, K. J. (2023). *Free agents*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691226231/free-agents

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Noble, R., & Noble, D. (2018). Harnessing stochasticity: How do organisms make choices?. *Chaos*, *28*(10), 1–7. https://doi.org/10.1063/1.5039668

Ostrom, R. G., & Walker, J. (1994). *Rules, games, and common-pool resources*. University of Michigan press.

Peters, U., Krauss, A., & Braganza, O. (2022). Generalization bias in science. *Cognitive Science*, *46*, 1–26. https://philpapers.org/rec/PETGBI

Prum, R. O. (2017). *The evolution of beauty: How Darwin's forgotten theory of mate choice shapes the animal world*. Anchor.

Pueyo, S. (2018). Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, *197*, 1731–1736. https://doi.org/10.1016/J.JCLEPRO.2016.12.138

Samuelsson, R., Price, S., & Jewitt, C. (2022). How young children's play is shaped through common iPad applications: A study of 2 and 4-5 year-olds. *Learning, Media and Technology*, 1–19. https://www.tandfonline.com/doi/full/10.1080/17439884.2022.2141252

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Strathern, M. (1997). "Improving ratings": Audit in the British University system. *European Review*, *55*(5), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:33.0.CO;2-4

Triggle, C. R., Macdonald, R., Triggle, D. J., & Grierson, D. (2022). Requiem for impact factors and high publication charges. *Accountability in Research*, *29*(3), 133–164. doi: 10.1080/08989621.2021.1909481

Wouters, P. (2020). The mismeasurement of quality and impact. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 67–76). MIT Press.

Zhuang, S., & Hadfield-Menell, D. (2021). Consequences of misaligned AI (arXiv:2102.03896). *arXiv*. http://arxiv.org/abs/2102.03896