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