- **Introduction**
  - Code of Conduct
  - Communication -- Miro and Zoom Chat
  - Introduce Juno Suárez
  - Follow Juno on Mastodon: @juno@hachyderm.io
- **Discussion led by Juno Suárez**
  - The paper is available at: https://www.researchgate.net/publication/371866602_Dead_rats_dopamine_performance_metrics_and_peacock_tails_proxy_failure_is_an_inherent_risk_in_goal-oriented_systems
- **Wrap**
  - Thank you Juno Suárez!
  - Continuing the discussion: discord
  - Next paper: Suggestions!
  - Follow #PapersInSystems on mastodon

https://en.wikipedia.org/wiki/Goodhart%27s_law

Niall Murphy's video skit for LFIconf on incidents managing by metric: https://www.youtube.com/watch?v=5MJEcntVK8E

highly recommend this history of Cold War rationality https://press.uchicago.edu/ucp/books/book/chicago/H/bo16160491.html

"When a measure becomes a target, it ceases to be a good measure"

- Goodhart's Law

The "Cobra Effect" is a similar illustrative story for Goodhart's Law: https://en.wikipedia.org/wiki/Perverse_incentive

Example	Regulator	Agent	Goal	Proxy	Failure	Constraints
Neuroscience	Various	Various	Health	Dopamine	Addiction	self, social
Corporation	Director	IC	Profit	KPIs	KPI hacking	intrinsic motivation
Software testing	Technical leaders	Engineers	Quality Code	Test coverage	Tests miss systemic failures	Other code and test standards
Counting Flies		Ambitious military officers	Ostensibly demonstrate adherence to military bureaucracy	Successfully completed form		
Public Safety	Diffuse - combination of political leadership, media, etc.	Criminal legal system & law enforcement actors	Community safety	Reported crime, arrests	To a large extent "crime" is a function of social conditions and pursuit of proxy can decrease public safety	Political counter-pressure, reform-minded legal system/law enforcement actors
"Lines of Code"	VP Eng	IC	"Productivity"	Lines		

https://excalidraw.com/#room=922d85b620c3e9c944c5,0BhEMCxCBqHqCnJB04KHfw

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.

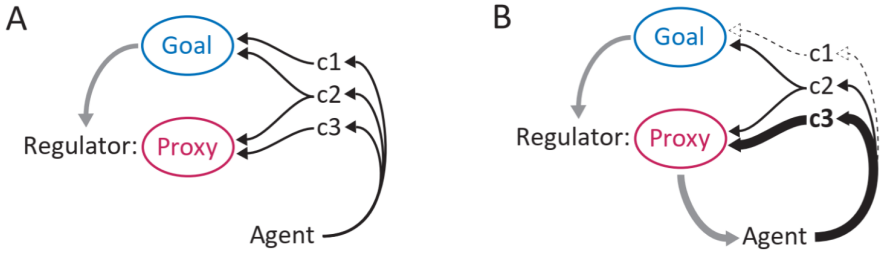
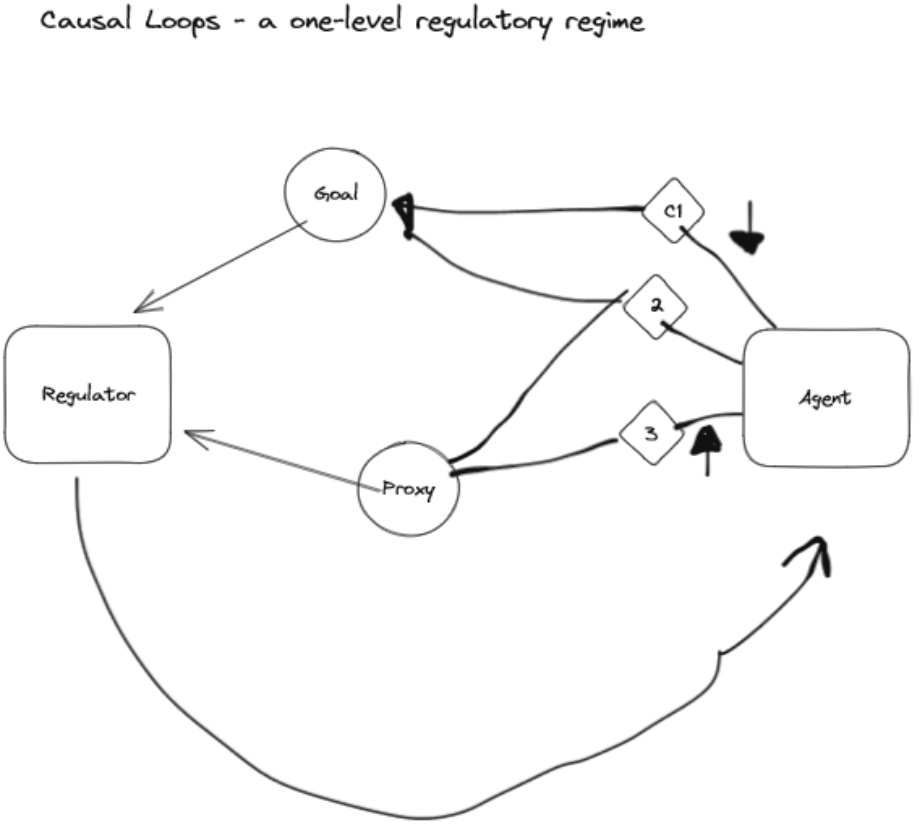


Figure 1. Regulator, Goal, Agent, Proxy, and their potential causal links. Proxy failure can occur when a regulator with a goal uses a proxy to incentivize/select agents. **A)** In complex causal networks the causes (arrows) of proxy and goal generally will not perfectly overlap. There may be proxy-independent causes of the goal (c1), goal-independent causes of the proxy (c3), as well 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

140525X23002753 Published online by Cambridge University Press

Who gets to decide if we're moving towards the goal?

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://excalidraw.com/#room=922d85b620c3e9c944c5,0BhEMCxCBqHqCnJB04KHfw

Similar story to the apocryphal story of the "cobra effect": "The term cobra effect was coined by economist Horst Siebert on the basis of an anecdote of an occurrence in India during British rule. The British government, concerned about the number of venomous cobras in Delhi, offered a bounty for every dead cobra. Initially, this was a successful strategy; large numbers of snakes were killed for the reward. Eventually, however, enterprising people began to breed cobras for the income. When the government became aware of this, the reward program was scrapped. When cobra breeders set their now-worthless snakes free, the wild cobra population further increased. This story is often cited as an example of Goodhart's Law or Campbell's Law." - https://en.wikipedia.org/wiki/Perverse_incentive

Lines of code is an easy software one to put there perhaps.

https://en.wikipedia.org/wiki/Hugh_Troy

Meaning, lines of code as proxy for productivity or progress of an system.
Body Mass Index as proxy for "healthiness".
https://garlanddavis.net/2018/08/11/the-fly-paper-report/

Similarly, defects fixed, rewards sloppier initial development.
“Days spent in an in-person office
Completed story points - pumping up the estimations.
LOC as proxy for IC productivity
when we get to the model itself, I've got some commentary/background on the Economics/GDP example that might enrich the model. no hurry to bring it up
Old maxim: the only good diff is a red diff
LOC provides perverse incentive for complex code
Lengthy code can also demonstrate disordered thinking.
I call that the "looking under the lighpost for the keys" problem. Using the easy measure, lines of code, to address a problem/goal that's really, really hard to find a good proxy
Knowledge that you’ll have to maintain it later
Peer review
Code review, pull requests
Process

In re the discussion with Shauna, I thought of Jacob Hacker’s idea of risk shift https://politicalscience.yale.edu/publications/great-risk-shift-new-economic-insecurity-and-decline-american-dream

https://excalidraw.com/#room=922d85b620c3e9c944c5,0BhEMCxCBqHqCnJB04KHfw

Is LOC just an example of a Corporatrion’s KPIs?

yeah, seems like it.

one we all relate to more easily than others :)

It used to compete with hours spent in the office

lol, lines of code/hour at home vs lines of code/hour at office

It’s easy to underestimate the chaos that agents can intentionally create.

I’m really good at malicious compliance, and other people are way better at it than I am.

and this is kind of pointing at malicious proxy creation to obfuscate the real intent. I have an example of that for that section on appropriation in this paper

I put this on the micro board, but Niall Murphy’s skit for the learning from incidents conference strikes me as very relevant: https://www.youtube.com/watch?v=5MJEcntVK8E

From another side, it becomes funny when management conceals proxies together with the goals, so that the proxies cannot be deduced. "Just continue what you're doing". What exactly? Why? Did someone experience this?

In my case there were no clear goals; or the goals that were, were moving.

Thinking of all the times people get told “you can’t move cards backwards on the kanban board because it will mess up the metrics”.

Oof, the history of colonization/settling deserves its own paper as an example of proxy failure.

Does this mean that OKR is not that sound?

https://cwodtke.medium.com/you-cant-handle-okrs-5465cf161e81

	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	Govt., Funders	(Campbell, 1979; Koretz, 2008; Nichols & 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	Govt./Military leadership	(Yankelovich, 1972)
	Cobra effect	Fewer cobras	Dead Cobras	Citizens	Government	(Siebert, 2001)
	Management **Indicatorism**	Profit/ firm value	KPI, Quarterly returns, …	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 & Garrabrant, 2018)
	Social media	e.g. entertainment	# of clicks/ time on platform	Content (e.g. videos)	Social media corporation	(Bessi et al., 2016; Faddoul, Chaslot, & 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 & Lippman, 2020; Braganza, 2020)
	Economics	Prosperity/ wellbeing	Profit/ GDP	Companies	Market	(Braganza, 2022; Kelly & Snower, 2021)
	Politics	Good governance	Votes/ popularity	Parties/ Politicians	Election	(Finan & 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 & horses)	Offspring quality	Chemical signal	Embryo	Parent	(McCoy & Haig, 2020)
	Embryo selection (plants)	Offspring quality	Chemical signal	Embryo	Parent	(Shaanker et al., 1988; Willson & Burley, 1983)
	Sexual selection	Mate fitness	Sexual signal	Displaying sex	Choosing sex	(Albo et al., 2011; Backwell et al., 2000; Funk & Tallamy, 2000; Gasparini et al., 2013)
	Runaway niche construction	Biological/ cultural fitness	Physical/ behavioural trait	Selected trait	Constructed niche	(Rendell et al., 2011)
	Neonate selection (marsupials)	Offspring quality	Speed to find teat	Neonate	Mother	Present paper
Neuroscience	Preference learning	Utility/ fitness	Reward signal (e.g., dopamine)	Preferences/ habits	Organism/ meta-cognition	Present paper
	Diet	Nutrition/ health	Sweetness/ saltiness reward	Food re-presentations		
	Addiction	Learning	Dopamine bursts	Plan/habit re-presentations		
	Exploration	Knowledge	Novelty related reward signal	Plan re-presentations		