Building Less Flawed Metrics

Manheim, David

26 November 2018

Building Less Flawed Metrics
Dodging Goodhart and Campbell's Laws

David Manheim

November 26, 2018

1 Abstract

Metrics are useful for measuring systems and motivating behaviors. Unfortunately, naive application of metrics to a system can distort the system in ways that undermine the original goal. The problem was noted independently by Campbell[Cam79] and Goodhart[Goo75], and in some forms it is not only common, but unavoidable due to the nature of metrics[MG18]. There are two distinct but interrelated problems that must be overcome in building better metrics; first, specifying metrics more closely related to the true goals, and second, preventing the recipients from gaming the difference between the reward system and the true goal. This paper describes several approaches to designing metrics, beginning with design considerations and processes, then discussing specific strategies including secrecy, randomization, diversification, and post-hoc specification. Finally, it will discuss important desiderata and the trade-offs involved in each approach.

2 What is the Problem, Exactly?

Metrics, key performance indicators (KPIs), targets, quantifiable goals, measurable results, and objective assessments are a few of the terms that get used to refer to the modern obsession with numerical and therefore seemingly scientific ways to understand human systems. These trends have led to improvements in business processes, in medicine, in public safety, and in both primary and higher education. In part as a result of this success, there have been highly publicized failures of the ever-more commonly applied paradigm. These occur when the measure isn't aligned well enough with the true goal, when the system promotes cheating, or when a formerly useful measure is applied despite underlying changes that make it no longer relevant.

Both Campbell[Cam79] and Goodhart[Goo75] identified an important failure mode for measurement, which was later paraphrased by Mary Strathern [Str97] as "When a measure becomes a target, it ceases to be a good measure." Campbell, who seems to have discovered the concept first[Rod17], was looking at social science and the way in which metrics distort behavior and lead participants in a system to attempt to exploit metrics. Goodhart, on the other hand, was an economist noting a structural breakdown

in inference about a system which occurs when rules change - a precursor to the now-famous Lucas critique in economics. The dynamics involved in these failures, however, are more complex than either discussed at the time, and several distinct failure modes and underlying dynamics have been identified[MG18], which can be simplified into a few cases.

2.1 Delineating the Problems

First, a metric that is currently statistically correlated with the goal will inevitably be less closely correlated once the metric is used, for example when conditioning on high values of the metric. As an intuitive example, height and basketball skill are correlated, but among the tallest people, it is unlikely that the best few basketball players are also the tallest. A similar problem occurs when a metric is correlated with an intermediate measure which itself correlates with a goal. For example, high school grades correlate with college success, and all else equal a student who takes easier classes in high school will receive higher grades - but taking easier classes will not lead to greater success in college. In both cases, simplification of the metric has important effects; easy to measure is rarely the same as important[Hub14].

Second, there are times when explicitly optimizing a system using a metric will change the system to make the metric invalid, as Goodhart noted. Students who sit near the front of the class typically get better grades, but if a teacher seats the worst-performing students in front, the relationship will likely disappear. An important example of this breakdown, and the one Campbell was referencing, is when the participants in a system explicitly react to the new rules. If the teacher in the previous example announced that instead of quizzes, they will assign a portion of the grade based on seat position, the new incentives will make the grades less useful for measuring learning.

Lastly, the discussions above make an implicit assumption shared by both Goodhart and Campbell, that the goal is coherent and understood. In some cases, however, the goal is incoherent. A simple example is a committee composed of individuals with differing values and goals. Because the individual goals can be incompatible, there may be no coherent way to assign a metric that achieves the different and incompatible goals. If the choice of metric is a compromise that doesn't address the fundamental conflict, the actual incentives chosen may be incoherent. A similar problem occurs when the desired outcomes are unclear. An example of both of these is the education system. The desired outcomes of education include life-satisfaction, fitness for the future job market, fostering the intellectual curiosity of students, or creating informed citizens. These are all long-term, and thus hard to measure or discuss concretely, are not often discussed by those setting priorities, and are often conflicting. Unsurprisingly, various intermediate metrics like GPA, even at the college level, or college completion, are poorly correlated with the desired long-term outcomes[Cap18] - and the difference is subject to gamification[Hes18]. This is unsurprising - the degree to which incoherent, conflicting, or poorly defined goals can be achieved is intrinsically limited. Worse, as Deresiewicz argues [Der15], imposing simplistic metrics distorts education in a way that defeats the original goals.

2.2 Addressing the Problem

The problem statements above seem to suggest solutions. These solutions are not always simple or practical, and as we will explore later, the approaches are viable and acceptable in different areas.

To address the problem of collapsing correlation, it seems possible to build metrics that more closely relate to the actual goal. In our first example, instead of using height as a proxy for basketball ability, we can use a weighted sum of height, athleticism, mastery of basketball skills, and experience. This will improve the model, but unless aclear causal model for basketball ability is found, it will be only a partial solution. In the second example, we can explicitly measure the relationship between student behavior like choosing easier classes and college success, instead of making the mistaken assumption that correlation is transitive. Unfortunately, investigating all the potential confounding interactions between high-school choices and college success (which itself must be measured in ways that are fallible,) is a much larger project, and it still does not ensure that causal mistakes would not allow other forms of collapse. For example, perhaps hours of studying is caused in large part by interest in academic subjects, which causes later success. Selecting students who participate in study groups that then get listed on college applications would seem to help, but involuntary or boring study sessions might be due to poor grades and disinterest in the subject, and anti-correlate with the actual cause of later success.

To address the second problem, of metrics distorting the system, we need a two pronged approach. The first prong requires insisting on metrics robust to changes, such as ones using models of the system that represent how the measured quantity relates to or affects the goal. In the example, if the causal relationship between seating and performance is understood, the chosen metrics will properly represent the determining factors of the relationship, such as student motivation and attention paid. The exercise of thinking through the causes will hopefully make it clear that re-arranging seats will have minimal effect. While these observations are sometimes obvious, discovering causal relationships is in general complex. The second prong is ensuring metrics are not being manipulated by the participants, or at least minimizing this manipulation - via secrecy, randomization, or post-hoc choice of metrics. For example, if students are unaware that grades will be assigned based on seat position instead of work done, their actions will less severely distort the metric.

Lastly, incoherence and debated goals can sometimes be addressed with structured discussions leading to increased clarity. In such situations compromise is often needed. Where clarity and compromise are possible, coherent goals are found that can (in the terminology of the late, great Herbert Simon) satisfice - that is, find solutions that are acceptable instead of optimal. Abandoning the search for an optimal solution or compromising on key goals may seem unfortunate, but the alternative of using incoherent metrics is often worse than doing nothing at all.

3 Important Desiderata Across Domains

The "Scientific Management" movement was an early proponent of reward systems similar to those seen in use in corporations today; profit sharing, per-task payments or bonuses, and merit-based pay[CP14]. In each case, the reward is tied to a metric. On the other hand, motivators are complex, and there are important trade-offs between the various positive and negative reward factors[Her68].

In addition to the operational challenges, there are various desiderata involved in actual decision-making around metrics that may be implicated. Metrics often benefit from immediacy, simplicity, transparency, various forms of fairness, and non-corruptibility. The exact trade-offs between various motivational factors are a matter of intense empirical focus, but stepping back from those discussion we can see that the desiderata mentioned are all implicated.

Immediacy is useful for ensuring feedback can be applied quickly, and participants can learn what is expected. For example, delayed rewards like end-of-year bonuses may be less effective motivators. Overly complex metrics may be less effective in motivating behavior, and impose costs on both the participants and the evaluators. Transparency is important for trust, may be a regulatory or legal requirement, and can avoid principle-agent problems. Secrecy also undermines perceptions of fairness, which can create issues of trust. Fairness is also important for legal and social reasons, and even if an unfair metric is able to accomplish the intended narrow goals, it can lead to longer term issues and undermine social trust. Corruption, of course, is a more direct attack on many of these desiderata, and either the perception or the reality of manipulation can do enough harm to more than outweigh any possible benefit from the use of a metric. More central to the problems of Goodhart and Campbell's laws, employees almost always analyze the system and are intentionally or unintentionally motivated to circumvent the intent to achieve the stated goals.

The use of rewards to motivate behavior is, of course, not limited to the domain of for-profit business, and the trade-offs in other domains can differ. Public policy often uses tax incentives, which have limited effectiveness due to complexity, non-immediacy, and suspicions of unfairness. In the measurement of autonomous vehicles, a recent report suggested that the measures must be "valid, feasible, reliable, and non-manipulatable," [FBBAK18] implicating many of these same concerns. Punishment systems have many of the same features - law enforcement is less effective when arbitrary, when the punishments are often avoided, or when the perpetrators of "crimes" find technical ways to avoid culpability. Prize competitions are an attempt to use motivators even more directly, but participation will be limited if potential recipients worry about unfair treatment or corruption. Lack of clarity about goals would be even more critical.

4 Strategies and Trade-offs

The first five strategies are ones involving the process of creating and considering the metric. These process-oriented methods are not reflected in the metric itself, but can

lead to better choices of metric. The next five are potential properties of the metric themselves, and the earlier processes can consider these methods when selecting metrics. The final strategy is considering the effectiveness of using concrete numerical metrics altogether, because not all problems can be effectively addressed using metrics. The strategies are of course not mutually exclusive, and they are often complementary, but the list is intended to be exhaustive.

4.1 Design Considerations

There is a common temptation, to find easy to measure outcomes instead of choosing based on how well a measure represents the goals[Hub14]. Unfortunately, this temptation is too-often yielded to in practice. There is a trade-off between ease of measurement and accuracy, but the choice should be made based on consideration of the options. In order to accomplish this, there are three general thought processes that may be useful in avoiding metric over-optimization failures.

Coherence. If the goals of a system are incoherent, or are poorly understood, it will be difficult for any metric to capture them. Incoherent (or under-specified) goals often lead to measuring whatever is most convenient to measure, instead of measuring something important to the process [Man16]. For example, it is easier to measure lines of code written by a programmer than it is to judge how well the code performs. In some cases, the metrics in place serve simply to justify the status quo, or to act as window dressing. Promotions in companies may in theory be based on metrics, but if managers can choose to apply the metrics selectively, this can serve as a mask for justifying decisions made on a different basis.

Structured Discussions and Compromise. In situations of deep uncertainty and conflicting goals there is often a need for discussion and compromise. While no compromise can achieve conflicting goals, deep exploration of problems can often lead to agreements that are better for all participants than the alternatives[RM01]. Unfortunately, these methods require extensive analysis and discussion, and are ill-suited to many smaller-scale problems.

Causal Forethought. Sometimes the metric measures something related to the intended goal with an unclear or non-causal relationship. If this is the case, a reward system using that metric can creates incentives that make the relationship between the metric and the goal disappear. For example, measuring attendance in class may increase attendance, but if the otherwise-absent attendees spend their time in class sleeping, or being disruptive, it is possible that nothing will be gained. A theory of change is helpful for clarifying these relationship and avoiding this class of error. (See Taplin and Clark's book[1] [TC12], for a clear introduction to theory of change.)

Pre-Gaming. If a metric is proposed, the exercise of imagining how it could be gamed, and building incentives aimed at forestalling gaming, can be useful. This idea is

[1] Available online here: `http://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/ToCBasics.pdf`

closely related to research about the effectiveness of such planning by Mitchell, Russo, and Pennington[MEP89], which Gary Klein later popularized as a "pre-mortem"[Kle07]. After identifying likely failure modes, it may be possible to improve the metric, or add explicit conditions to the rewards to thwart the failure modes that were discovered. Despite the desire to restrain gaming, however, care should be taken to ensure that the metric does dictate exact methods, which can stifle innovative for accomplishing the overall goal. For example, measuring hours of classroom time spent by a teacher may discourage time spent on lesson planning, peer consultation, and other activities that improve effectiveness of the time spent in class. Explictly requiring each of those specific activities to account for the potential failure, however, removes discretion that allows teachers to pick the activities that are most beneficial in their case.

Monitoring Behaviors. Even when well designed and initially effective, metrics have a tendency to go awry over time as systems and behaviors change. Explicitly setting checkpoints and reviews for metrics may be useful for ensuring that these systemic drifts are limited in scope. This is especially useful when it is easy it is easy to detect behaviors which effectively cheat[2]. For example, metrics often promote a short term intermediate goal, like sales of a certain product, or short term ad-revenue. Incentives may start encouraging overzealous sales activities, or placement of ads that interfere with user happiness or engagement, in each case potentially preventing longer-term growth. Overzealous sales activities would be visible in lower repeat sales or reduced customer satisfaction, maknig detecting this failure relatively easy. Designing perfectly coherent metrics aligned with goals for the system overall may be infeasible, but monitoring behaviors that metrics incentify can detect or prevent larger distortions and later systemic failures.

4.2 Metric Features

Diversification. If a single metric does not align perfectly with the goal, introducing additional metrics, even if they are individually less well correlated to the goal than the first, can sometimes improve the system overall. Recalling the example above, the choice of the best basketball players is better predicted by a combination of metrics than any single one. In a similar way, it is often the case that multiple different metrics are better aligned with the true goal than any single metric. Because the different metrics typically require different behaviors, and they will be to some extent in tension with one another, they are likely to make gaming harder.

Secret Metrics. If the metric is not known to participants, they cannot game it. The existence of an un-revealed metric can still incentivise participants to achieve the goals they think most likely to be measured or rewarded, and to the extent that they understand the goal but not the metric, this will align incentives while preventing or at least hindering manipulation.

Post-Hoc Specification. If the metric is chosen after all actions are taken, participants view the metric as secret, but because the order of choices is reversed, attempted

[2]I am grateful to Davide Balzarotti for this insight.

gaming of the metric can be punished. This may be perceived as allowing unfair discretion, or may lead to actual corruption.

Randomization. Even if a metric is known beforehand, if the weights on components or the relative rewards are uncertain, gaming the metric may become less worthwhile. When done correctly, many forms of randomization can also allow much clearer evaluation of success, which is especially useful for monitoring the usefulness of the metric or reward system.

Soft Metrics. Human judgment, peer evaluation, and other techniques may be able to reduce gaming specific to metrics. Metrics are often seen as a way to avoid subjectivity, but a combination of metrics and human judgment may be able to capture the best of both worlds.

Limiting Metrics. Failures are often the result of too much pressure on the optimization. By using metrics to set a standard or provide a limited incentive instead of a presenting value to maximize, the overoptimization pressure can sometimes be mitigated.

Abandoning Measurement. Sometimes, the value of better incentivising participants and the potential for perverse incentives issues make it worthwhile to be wary of what Muller refers to as metric fixation.[Mul18] As he suggest, sometimes the best solution is to do nothing - or at least nothing involving measurement.

4.3 Desiderata vs. Strategies

The degree to which a strategy fulfills various desiderata is important, and the importance of a desiderata in a given domain can be weighed against the importance of preventing gaming and the effectiveness of a strategy. We can therefore define roughly what is meant by the desiderata, and note where there are obvious advantages or conflicts that should be considered.

Immediacy: Can the metric be computed in real time? Does it provide feedback rapidly enough to align incentives?
Simplicity: Is the metric difficult to understand? Will participants understand it well enough for it to influence their behavior? Are the implications understood?
Fairness: Is the metric commensurate to actual goals? Does the metric provide disproportionate benefit to some groups? Do behaviors that get influenced by the metric impose costs elsewhere in the system?
Non-Corruptibility: Can the system be used by a party providing incentives to cheat? Does the metric introduce unfair information asymmetries?

	Immediacy	Simplicity	Fairness	Non-corruptibility
Considering Coherence		#	+	+
Causal Analysis		-		
Structured Compromise		-	+	
Pre-Gaming			+	+
Diversification		-	+	
Secret Metrics	-		#	-
Post-Hoc Specification	-			-
Randomization	#	-	-	+
Soft Metrics	#	#		
Limiting Metrics		-	+	
Abandoning Measurement	+	+	-	-

The table indicates which desiderata are likely affected by which strategy. Positive effects on each desideratum are indicated with a plus, while negative ones are indicated with a minus. Complex interactions are complex are noted with a hash, and these are sometimes positive and sometimes negative.

5 Considering Applications in Practice

Not all strategies are appropriate in all domains, and implementation is critically dependent on factors specific to a given system and the relevant actors. Still, systems chosen by public authorities face a higher burden for fairness and non-corruptibility, while those implemented in private business often require more immediacy. Incentives intended to motivate non-experts benefit more if they are simpler and easily understood, and those that impact people or organizations which must participate in a system, such as employees, or those that involve high reward, may need to be more game-proof.

The variety of concerns that exist make illustrating all potential issues infeasible, but it is worth considering each of the metric features and seeing how they can help with the design of better metrics, and how they might impact the various desiderata.

5.1 Diversification

Metrics which amalgamate multiple simple measures are often useful when individual measures are insufficient. In addition, when a metric includes only some parts of a goal, it implicitly pushes emphasis away from the others. If reading and arithmetic are each 50% of the measured outcomes from school, it means that science, art, and physical education are all 0%. Because the easy to measure parts of a system are quickly accounted for and optimized for, even rudimentary or obviously biased measures of the remaining outcomes can offer significant marginal value.[Hub14]. We already measure accomplishments in math and reading, so adding measures of time spent in arts classes will at least mitigate the pressure to remove those classes completely - and by doing so, lose important longer term benefits that are more difficult to measure for short-term evaluation[Hes18].

Disaggregating aggregate metrics can be useful in disambiguating problems caused by

Simpson's paradox. Comparing subgroup outcomes directly can reduce the incoherence of comparing aggregate outcomes, which is sometimes important. For example, Leibowitz and Kelley show examples where different sub-population sizes can make ranked education outcomes severly reverse direction. Once the success of subgroups is considered, diverse areas which perform worse in aggregate are found to better serve every sub-population, making the aggregate metric not only incomplete, but incoherent.[LK18]

Diverse, disaggregated, or compound metrics can also mitigate problems with other types of incoherence, such as disagreement or lack of causal understanding. This is because a scattershot approach will tend to limit the degree to which any one measure influences the system. Designers with conflicting goals can choose measures that assist with each, and the combination may be an acceptable compromise. Similarly, if the causal relationships are unclear, targeting multiple different parts of the system may constrain the amount by which the system is changed due to the new incentives. In either of these cases, however, the metrics are unlikely to be coherent. Keeping metrics disaggregated can make it hard to compare or incentivise results, but any method of combining conflicting or varied measures will make the overall system more complex. Such complex and incoherent metrics are also much less effective at motivating desired behavior, since it will be harder to identify how to target the compound metric. The complexity and incoherence of the metrics can sometimes reduce the degree to which participants can game metrics, but simultaneously make it harder for the designers to identify ways that participants may find to game the system.

When goals are complex but cannot be directly measured, measures of various components or correlated outcomes can be used. This may make the goal easier to achieve, since it replaces an unclear target with clear sub-targets, but it may also make it harder for participants to decide what they should focus on. As in the previous case, gaming of metrics will be harder, but so will identifying how it will occur and how to prevent it.

5.2 Secret Metrics

When qualitative goals are understood, keeping participants from knowing the details of the measurement system will limit the degree to which they can exploit the system. This requires some conception of the goal independent of the metric. In the worst case, the awardees don't understand the goal at all, and they will not be motivated by the seemingly-arbitrary rewards.

This is an effective strategy for preventing gaming, especially when pre-gaming methods discover important vulnerabilities of the various metrics that are hard to avoid. This works well if the metrics can be gathered without informing participants, and where the metrics that would be used are not obvious. The strategy will be less effective at preventing gaming if they can guess or infer the metric which will be used. Similarly, if the data collection to support evaulation of the metric is visible to participants, such as requiring them to take a test or gather specific data, it will be harder to hide.

Unfortunately, secrecy is prone to degrade over time as rewards are received and people can infer what is being evaluated. If a metric must be used repeatedly or in real time, it will be difficult to keep participants unaware of the details of the system.

Similarly, if managers or regulators who implement the system are themselves being judged on the basis of the measured results, or they can be induced by participants to divulge information, they may intentionally degrade the secrecy needed. For this reason, secret metrics are more helpful if used one time then changed, as occurs when new tests are written for students each year - and as that case illustrates, knowledge of the types of questions commonly asked can still confer unfair advantages.

5.3 Post-Hoc Specification

When results are seen and analyzed before the metric is chosen, there are a variety of ways to prevent gaming while preserving the transparency of the rewards.

Designing measures completely post-hoc often involves justifying intuition or decisions already made. To avoid this, post-hoc specification should be limited to only include some parts of the metric. For example, the weights on various measures may be chosen after all activities have finished, or certain measures may be discarded based on analysis of the outcomes. If this process is known to participants beforehand, the potential for metrics to be discarded or given low weights can serve as an incentive not to game them.

The first, and most significant disadvantage for such post-hoc decisions is unfairness, both actual and perceived. Transparency in the process for the post-hoc selection can mitigate this problem, as can ensuring that the decision is made by a party that is not directly involved. The second significant disadvantage is that the feedback and reinforcement is delayed, which can significantly reduce the effectiveness of a reward system. A key advantage is that the post-hoc specification can keep the measures simple and easy to understand.

5.4 Randomization

Randomization can be used to choose between different proposed metrics when there is disagreement, or can be used within the metric itself. Choosing metrics via chance may avoid difficult compromise that leads to incoherent results. Allowing part of a metric or incentive to be determined by chance can be useful for preventing exploitation. Like secrecy and post-hoc specification, randomization reduces the direct connection between behaviors and metrics, which has some of the same positive and negative impacts.

To the extent that the weights and rewards are randomized instead of chosen intentionally, the incentives will be less well aligned with the actual goal. The uncertainty may also be perceived as adding significant and hard to understand complexity, and reduce motivation to achieve goals. On the other hand, exploitation is similarly less rewarding. Randomization can also be perceived as unfair, either because it rewards individuals differently, or because it rewards factors in a way not proportionate to importance.

Randomization works particularly well in combination with other methods. For instance, the randomization of the outcomes of a metric based on diverse inputs can assign random weights to already-known components. Similarly, it can be used to remove concerns about corruption for post-hoc specification, by pre-specifying the randomization

to be performed at the end of a time period. If used beforehand to assign different metrics or different weights on metrics to different groups, it can also be valuable for analyzing the outcomes from using various metrics and incentive systems.

5.5 Soft Metrics

Metrics can include quantitative but subjective evaluations. These soft metrics are often able to avoid certain pitfalls of focusing on easily measured quantified values. For example, peer ratings by programmers will not reward behaviors that help achieve measurable results like rapid but sloppy development at a high cost to the overall goals and maintainability of a system. Such measures have their own potential for exploitation, where participants game the system via currying favor, "sucking up," or taking measures to appear more productive than the reality.

Such soft metrics can be done routinely, which has the advantage of providing feedback rapidly, but the cost of measurement can be high, if participants need to routinely spend otherwise productive time doing evaluations. They can also be perceived as unfair, and this can also lead to fighting or backstabbing - especially if the rewards are zero-sum.

5.6 Limiting Metrics

Metrics do not need to be maximized to be effective. If the metrics is used as a minimum for some incentive, the overoptimization may dissapear. By replacing optimization with what Simon terms satisficing[Sim47], many of these problems can be avoided. For example, bonuses for salespeople who hit sales number targets is less likely to lead to overly competitive employee dynamics, where employees try to "steal" credit, or aleinate customers with overly aggressive tactics.

This strategy is not always appropriate. Steven Shorrock noted that "when you put a limit on a measure, if that measure relates to efficiency, the limit will be used as a target."[Man18] His original example was of flight duty times, where a regulation limiting the maximum number of duty-hours that airlines crews can work led to use of that minimum as a target for airlines. Now that they must measure crew-duty times, airlines try to ensure their employees are as close to the limit as is possible. By introducing this new measure, it is possible crews are now more overworked than they were without it.

Satisficing can also allow complacency once targets are reached. Climate legislation limiting total emissions have failed because they were not ambitious enough, and "the shortcomings identified... are inherent to crediting mechanisms in general" [CHF+16]. That report found, as one important shortcoming, that transferrable emissions credits were worthless in part because there were too many credits that were being generated effectively for free. This was made worse because of the ability to transfer the credits from countries that exceeded the goal to places where the goal was not met. Because no further incentive was in place once targets were met, there was no need to embark on more ambitious projects. In such a case, structuring the incentive differently might have been more effective. For instance, a moderately-sized tax on emissions could provide

incentive to do some amount of mitigation without providing a potentially unlimited incentive to artificially game the system the way refundable tax credits might.

5.7 Abandoning Metrics

There are situations where measuring outcomes is too expensive to be justified by the potential improvement that it could create. This occurs when the complexity needed to correctly represent the system can require a business structure that is unreasonably or inefficiently complex[PP16]. In other cases, the measurement system is likely to lead to distorted incentives rather than the initial goal. The aphorism is correct that what isn't measured isn't managed - but when choosing between not managing part of a system by not measuring it, or measuring it in a way that makes it worse, the answer should be clear.

The negative impacts of poorly designed metrics are felt by multiple parties, not only those who the metrics are intended to help. Obviously, the people who promote the metrics would prefer if their actual goal were pursued. Anyone who attempts to target the ultimate goals of the system and ignore the perverse incentives are implicitly punished for not playing these games, and would prefer better metrics that reward their efforts. The people who do adopt strategies to exploit the perverse incentives may benefit directly, but would often be happier not to be forced to play the game of understanding and exploiting complex, changing, and often harmful systems. Their exploitation of metrics has impacts well beyond their satisfaction since the economic waste and negative externalities created by exploiting poorly designed metrics is huge.

Choosing not to manage a system is a decision that should not be made lightly - especially not before seriously considering whether an alternative measurement might be useful. On the other hand, putting in place a mediocre measurement system prematurely is often far worse. Until serious consideration has been given to the processes and alternatives identified above, it may be better to wait, or to abandon measurement, rather than deploy a system that will be ineffective or worse. As Muller puts it, "sometimes, recognizing the limits of the possible is the beginning of wisdom. Not all problems are soluble, and even fewer are soluble by metrics."[Mul18] These limits are particularly relevant if participants will be drawn to the explicit rewards that are less well suited to accomplishing the goal than those who would participate regardless. The limitations are also critical if participants feel discouraged by the extrinsic motivation and measurement, especially in domains where intrinsic motivation is primary. This is supported by the empirical work by Rasul et al. showing that autonomy, which is incompatible with extensive measurement and accountability systems, is more effective for civil service[RR17, RRW17, RRW18].

6 Conclusion

Despite the intrinsic limitations of metrics, the frequent use of poorly thought-out and badly constructed metrics do not imply that metrics are doomed to eventually fail, or

that they should not be used because they will be exploited. Instead, forethought and consideration of the problems with metrics is often worthwhile. This process starts by identifying and agreeing on coherent goals, then considering both what leads to the goals, and what parts of the system can be measured. After identifying measurable parts of the system, and considering how participant behavior might exploit the measurement methods or the measured outcomes, measures can be constructed. The construction of these metrics to avoid exploitation may involve multiple diverse measures, secret metrics, intentional reliance on post-hoc specification of details, and randomization. This may also include decisions about where subjective measurements are important, and consideration whether measurement will be beneficial. In building the metrics and decising whether to implement them, attention should be paid to various important factors in the system, including immediacy of feedback, simplicity and understandability of the measurement system, fairness, and the potential for both actual and appearance of corruption in the metric and reward system.

Metric design is an engineering problem, and good solutions involve both science and art. Following these guidelines will not make metrics unexploitable, nor will it keep everyone happy with the results of a process. This is true of metrics used for employees, metrics used for monitoring systems, and even metrics used within machine learning algorithms - in each case, poorly designed metrics will be exploited. Occasionally, the suggested process will lead to investigation of potential improvements or strategies that are ultimately decided against. Despite this, it is a vast improvement on the too-common strategy of using whatever metric seems at first glance to be useful, or deploying metrics without considering what they in fact promote. Putting in the effort to build elegant and efficient solutions won't fix every problem, but it will lead to less flawed metrics and better results overall.

Acknowledgements

I am grateful to Abram Demski and Scott Garrabrant for prompting my initial interest in Goodhart's Law, and Venkatesh Rao for allowing me to explore the ideas on his "Ribbonfarm" blog in a series of guest posts. Conversations with other Ribbonfarm auithors and readers, as well as twitter conversations about these ideas have been very helpful. I am especially grateful for the twitter-based insights provided by (in no particular order,) Steve Shorrock, Sam Gardner, James Pitt, Simon DeDeo, Thomas Dullien, Scot Hazeu, Alex Holcombe, Tiago Forte, Daniel Bilar, Nick Szabo, and I am sure others who I am unfortunately omitting. I'd also like to thank Paul Davis for insight into complex policy systems generally, and useful feedback on the paper, as well as Osonde Osoba for early conversations and feedback about how randomization and secrecy would impact metrics. Lastly, I would like to thank the Machine Intelligence Reserach Institute for prompting, and the Berkeley Existential Risk Initiative for funding, my work on metric-alignment for artificial intelligence in multi-agent systems, which has spurred much of my thinking on how the same dynamics affect human systems.

References

[Cam79] Donald T Campbell. Assessing the impact of planned social change. *Evaluation and program planning*, 2(1):67–90, 1979.

[Cap18] Bryan Caplan. *The case against education. Why the education system is a waste of time and money.* Princeton University Press, 2018.

[CHF+16] Martin Cames, Ralph O Harthan, Jürg Füssler, M Lazarus, C Lee, P Erickson, and Randall Spalding-Fecher. How additional is the clean development mechanism. *Analysis of application of current tools and proposed alternatives. Oeko-Institut EV CLIMA. B*, 3, 2016.

[CP14] Helene L Caudill and Constance D Porter. An Historical Perspective of Reward Systems: Lessons Learned from the Scientific Management Era. *International Journal of Human Resource Studies; Vol 4, No 4 (2014)DO - 10.5296/ijhrs.v4i4.6605*, dec 2014.

[Der15] William Deresiewicz. *Excellent sheep: The miseducation of the American elite and the way to a meaningful life.* Free Press, 2015.

[FBBAK18] Laura Fraade-Blanar, Marjory S Blumenthal, James M Anderson, and Nidhi Kalra. Measuring Automated Vehicle Safety. 2018.

[Goo75] Charles A E Goodhart. Problems of monetary management: the UK experience. In *Papers in Monetary Economics*. Reserve Bank of Australia, 1975.

[Her68] Frederick Herzberg. One more time: How do you motivate employees, 1968.

[Hes18] Frederick Hess. Straight Up Conversation: Scholar Jay Greene on the Importance of Field Trips. *Education Week*, sep 2018.

[Hub14] Douglas W. Hubbard. How to Measure Anything: Finding the Value of Intangibles in Business. *LIVRO*, 2014.

[Kle07] Gary Klein. Performing a project premortem. *Harvard Business Review*, 85(9):18–19, 2007.

[LK18] Stan Liebowitz and Matthew L. Kelly. Everything You Know About State Education Rankings Is Wrong: Minds and dollars are a terrible thing to waste. *Reason*, nov 2018.

[Man16] David Manheim. Overpowered Metrics Eat Underspecified Goals, 2016.

[Man18] David Manheim. Shorrock's Law of Limits. *Medium.com*, may 2018.

[MEP89] Deborah J Mitchell, J Edward Russo, and Nancy Pennington. Back to the future: Temporal perspective in the explanation of events. *Journal of Behavioral Decision Making*, 2(1):25–38, 1989.

[MG18] David Manheim and Scott Garrabrant. Categorizing Variants of Goodhart's Law. pages 1–10, 2018.

[Mul18] Jerry Z Muller. *The tyranny of metrics*. Princeton University Press, 2018.

[PP16] Konstantinos Poulis and Efthimios Poulis. Problematizing fit and survival: transforming the law of requisite variety through complexity misalignment. *Academy of Management Review*, 41(3):503–527, 2016.

[RM01] Jonathan Rosenhead and John Mingers. *Rational analysis for a problematic world revisited*. Number 2nd. John Wiley and Sons, 2001.

[Rod17] Jeffery Rodamar. There ought to be a law! Campbell v. Goodhart. 2017.

[RR17] Imran Rasul and Daniel Rogger. Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. *The Economic Journal*, 128(608):413–446, 2017.

[RRW17] I Rasul, D Rogger, and M Williams. Management and bureaucratic effectiveness: A scientific replication. 2017.

[RRW18] Imran Rasul, Daniel Rogger, and Martin J. Williams. Autonomy, incentives, and the effectiveness of bureaucrats. *VoxDev*, 2018.

[Sim47] Herbert A Simon. *Administrative behavior; a study of decision-making processes in administrative organization.* Macmillan, 1947.

[Str97] Marilyn Strathern. 'Improving ratings': audit in the British University system. *European Review*, 1997.

[TC12] Dana H Taplin and Hèlene Clark. *Theory of change basics: A primer on theory of change*. 2012.