Rationally Speaking #240: David Manheim on “Goodhart’s Law and how metrics fail”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and my guest today is David Manheim.

David is a decision theorist with a PhD in public policy from Pardee Rand Graduate School. And one of the topics that David has studied and written a lot about over the years, in blog posts and academic articles alike, is a principle called Goodhart's law.

It's in that small set of deceptively simple principles that once you understand it, kind of explains so much of what's wrong with the world. So Goodhart's law, you might have heard it stated as “When a measure becomes a target, it ceases to be a good measure.”

We're going to talk today about what that means, how Goodhart's law shows up and and kind of the dynamics of how it works. So David, welcome to Rationally Speaking.

David Manheim: Thanks. I'm excited to be here.

Julia Galef: I'm curious how you got interested in Goodhart's law in the first place, and specifically whether it was more like, seeing how consequential this law is to education and healthcare and policy and business and things like that, in the real world? Versus the kind of mathematician’s, “Wow, what an intellectually interesting set of dynamics for me to puzzle over”?

David Manheim: So it's interesting. It was kind of a weird path. I had talked about it a little bit at a Less Wrong meetup in Los Angeles with a couple of people who are now at MIRI.

And then I was writing a blog post about how corporations figure out how they organize themselves. And a bunch of people commented that corporations should be able to do this really easily. They'll just set targets for what it is that different groups or business units should do and tell the business units to do that. And the business units can kind of go off on their own and just have the marketing group optimize to get as many people as possible to click on the website. And then have the sales people optimize to sell as many products as they can.

And it turns out that this really, really doesn't work, if you actually get people to work on really narrowly defined targets. As the law says, things start going wrong.

Julia Galef: So that was the context in which you first started getting interested in Goodhart's law, was organizational theory?

David Manheim: Yeah. And that was much more closely related to my work in grad school on bureaucracy, and how it is that organizations work, than it was to what I later ended up thinking a lot more about, which was how that matters for AI.

Julia Galef: Great. Yeah. And I want to talk about both the organizational theory context and the AI context. But let's first just kind of get more of a handle on what Goodhart's law is and how it works.

Maybe the prototypical example, like when people write blog posts about Goodhart's law, the illustrative example they start with is a story that is probably apocryphal from the former Soviet Union.

David Manheim: Ah, yes.

Julia Galef: Where the government had this measure of factory performance that they would use to incentivize factory owners. And it was based on the number of nails produced. So yeah, factories that were supposed to produce nails. The managers were judged based on how many nails they produced. So as a result, of course the factories produced millions of nails that were incredibly small and not actually useful for anything.

So then the government was like, okay, nevermind. The thing we're going to judge you on is the weight of the nails that are produced.

And so then of course the factories were like, great. And then they made just a small number of extremely large heavy nails that were also useless for anything.

And the point being that any kind of simple way that you define how people are being judged or graded or rewarded or just evaluated in any way, any simple metric like that is kind of easily gamed. Like the nail metric.

Would you consider that a central example of Goodhart's law, or is there a different one that you think is a better illustration?

David Manheim: So there are a couple of different ways that Goodhart's law manifests. The central dynamic of people, I guess “munchkin-ing” rules.

Julia Galef: Where munchkining is --

David Manheim: Trying to figure out how to use the rules that are there to do things, as well as they can, to do what they want to do. And kind of ignore the point of the rule.

Julia Galef: The spirit of the law. Yeah.

David Manheim: Right. So there are a couple of places where that happens. I actually think the clearest example of people trying to beat the rules and ignore what it is that's happening is a scandal that has happened a couple of different times for exactly the same reason every time, which is:

Teachers going in and changing students' answers on tests for standardized tests. They're just directly changing what it is that the result is, so that they look better. And it's not a sophisticated game that they're playing... They're just changing things so that they do better.

And you have to put a lot of things in place to make sure that you can trust the numbers that come out of a system where you're paying people or motivating them. Or in the Soviet case, threatening to throw them in the Gulag if they don't manage to do what you want them to. It's really hard to get them not to play games then.

Julia Galef: Right. I mean that, the case of teachers actively going in and altering students' answers is almost a less interesting example to me than other stuff that happens, to try to boost standardized test scores.

It might've been in one of your blog posts -- you wrote about teachers kind of half-consciously “teaching to the test” at the expense of teaching the underlying principles involved. Like telling students, “Okay, just plug in -- on a multiple choice question -- just plug in each of the possible answers into the equation to see which one makes it come out right. And that's how you know what the right answer is.”

Which is a way to get the right answer, but it doesn't help you understand the algebra involved.

David Manheim: And it doesn't generalize to anything other than a multiple choice test where you're given four answers you can plug in really quick. If it had been a fill in the blank test, you'd be stuck. You can't just plug in numbers until you find one that works. That might be a different strategy. Figuring out how to find closer approximations is a

reasonable strategy. But just plugging in the numbers you're given isn't.

So, yeah, I've talked about that a couple of times. Teaching to the test is one of these things that I think illustrates a slightly different part of the dynamics, which is... Eliezer Yudkowsky talks about, in a blog post he wrote a bunch of years ago, "lost purposes." Where he says, you have an organization that starts out and says, "Look, we need to do education. And education means people need to be able to do this thing – 'this thing' being in this case mathematics.

David Manheim: They need to be able to do algebra."

So how is it that you teach somebody to do algebra? You have to cover these lessons. So what are the teachers told? "Here's the set of lessons that you need to cover."

So teachers go through dutifully and cover the lessons that they are told to cover.

Does that necessarily align with making sure that all of the students actually understand the subject? No, definitely not. Most students that struggle with math, I would say in high school, probably need somebody to spend a bunch of time with them working on how fractions work, rather than plugging through more algebra.

And so what happens is the goal, which is making sure people know how to do things with math, has been lost to the purpose or to the narrow set of things that they've been told to do.

Well, the narrow set of things that they've been told to do are prepare people for standardized tests. Teachers hate this universally. If you talk to teachers, they say, "We hate that we're teaching kids, we're spending, a half dozen classes just doing SAT prep in our math class. Like, we could be teaching them something then."

They're not doing this because they want to get away with something, usually. What they're usually doing is following the rules that they've been given, to achieve targets that have been set, because somebody lost track of the fact that what it is we're actually trying to do is graduate students who know how to do this.

Julia Galef: So that is a really interesting question of what exactly is going wrong there.

So there's... it seems like one category of Goodhart's law is when people genuinely just have different incentives. Like in a telemarketing company, for example, maybe the management cares about the profits of the company. Like, upper management cares whether the company is doing well over the long run or not. Or at least the near term.

But the workers themselves do not care. And so the management might come up with a rule like "Your bonus is based on the number of calls you complete in a night," or something. And so the workers now just do calls really quickly, but they don't care about the quality of the call. So their sales actually go down.

And the workers may know that that's what's happening, but they don't really care because all they care about is getting their bonus. They don't really care about the company as a whole.

So that would be a case of just misaligned incentives. Or, yeah, incentives that are at odds --

David Manheim: I was going to say, there's an important kind of subcategory of what happens there, which is you usually have different business units inside a company that are legitimately trying to work on different parts of the problem, but end up in a situation where the people who are doing lead generation hand off leads that aren't actually going to make much money. And the salespeople are trying to figure out how to maximize the total dollar value of sales instead of the profits.

And senior management is sitting there looking at it, going, well, how do we get all of these different groups to work together? And the answer is, getting people to work together is hard. Running large organizations is hard.

And a simple solution is to give them clearly defined goals. Sometimes it's even the best solution, even though it falls prey to this failure mode.

Julia Galef: Right. But would you agree that's still a separate category, from the category where the different people in the system just genuinely care about different things, and they don't have the same end goal? The thing you're talking about here is just, "We all have the same end goal, but it's really hard to coordinate together to achieve that goal."

David Manheim: So inside of organizations... this is something that's more general than Goodhart's law, but I think critical for understanding it a little

bit better. Which is, most of the time people's incentives have a lot to do with context and only a little bit to do with, kind of, "management dictates." Usually people are doing the things that they do because this is what they're being told to do. This is what their manager wants them to do. This is what all of the people around them are doing.

So most of the time, it's easy to conceptualize things as "this is a straightforward principal-agent problem, where the principal wants X and the agent wants Y, and you need to figure out some way to get them aligned with one another."

Julia Galef: Where the "principal" is a person who sets the goals or the orders, right, and the "agent" is the one carrying out those orders?

David Manheim: Right. So in economics, this is a big topic of discussion and there's tons of work on this, that assumes that there are these nicely defined objectives that you're trying to maximize.

And that gets to, I think, another part of the discussion, which is -- the reason why this is hard is because we don't have a really good idea about how to define what our goals are.

So in a company, you can say, "We want to maximize profit." And even that isn't really true. We want to maximize profits subject to not having PR fiascos and having our executives thrown in jail for violating laws. There are a lot of things that you're trying to do.

Julia Galef: Right. And not in a way that will cause us to burn out in six months, or...

David Manheim: Right. So these are all things that actually matter. How do you operationalize all of those? So, even ignoring those constraints, how do you operationalize "maximizing profit" in a way that tells individuals in the company what they're actually supposed to do?

You tell the guy sweeping the floor, "By the way, sweep the floor in a way that maximizes profits"? That doesn't mean anything.

Julia Galef: Right. Right.

David Manheim: You do have to operationalize everything. And if you don't have a clear mental model, an actual model of how it is that everything relates to one another... which is hard to do. It's hard to figure out what it is you actually want. So it's really hard to figure out how to set goals that accomplish it.

Julia Galef: Can I give you a few examples of phenomena, and you can tell me if they count as Goodhart's Law or not? This is one of my favorite ways to try to understand the boundaries of the definition of something -- is just throw examples at someone who understands it, and have them tell me yes or no, and why not, if not.

So one example that actually came up on Twitter last year was a news story that I shared, about an aquarium that tried to train the dolphins in their aquarium to clean up the litter that people would toss into the pool. And so what they would do is they would reward the dolphins by giving them a fish if the dolphins brought them a piece of litter. Or like, a dead seagull.

And then one dolphin started tearing pieces of litter into smaller pieces -- because they were being rewarded based on the number of pieces of litter. So then they would trade each torn piece of litter in and get a fish for each one.

And then another dolphin started stockpiling fish. So they would get fish as a reward, but they wouldn't eat it right away. And they would stockpile the fish to then lure seagulls into the pool, and kill them. And then trade the dead seagull in for more fish.

Which was just so ingenious. I'm a little scared of dolphins. Like if they had opposable thumbs, I would be genuinely terrified.

But anyways, there was a debate in that thread about whether that is an example of Goodhart's Law. And I think you said it wasn't, so why not?

David Manheim: So there are a couple of pieces of this that are going on, and it depends on exactly where you want to draw your lines. And some of that matters and some of it doesn't.

But the key thing that happens when we're talking about people falling prey to Goodhart's Law, organizations falling prey to Goodhart's Law, is that somebody mistakes the metric for the goal. So in organizations, what that means is that at some point, the purpose was lost. Here, I don't know if the purpose was lost. I think it was just somebody -- you know, “somebody”, a dolphin -- came up with a clever way to beat the system.

Julia Galef: I think they count as somebody. I think any creature that can do something that clever counts as “somebody”.

David Manheim: I think that's fair. So I would say it has a lot of aspects of the principal-agent dynamic where they're not doing the thing that you

wanted them to do because you're paying them, in fish, to hand you a number of things.

David Manheim: So they're making small nails, but it's not due to the fact that there was some confusion at some point about what the goal was or some... There was never a point where somebody said, "Oh, well, the only way we have to measure this is this." It's just that the trainers found this to be the easiest way to implement the system.

Julia Galef: Right, the purpose wasn't lost. It was ignored.

David Manheim: Right. And sometimes that's, as I said about companies, sometimes that's fine. I don't know if it's such a horrible thing for the clever dolphin to be ripping up the garbage and handing them multiple pieces to get more fish.

Julia Galef: It makes it a less efficient metric, but it still works, basically?

David Manheim: Right. At the point where it's killing seagulls --that's definitely a more problematic failure mode.

Julia Galef: Right, that's a good distinction. Okay. What about another example to throw at you... What about the user engagement metrics that social media companies like Facebook use, which end up furthering the spread of sensationalist or even false news, because that's the kind of stuff that causes higher user engagement?

Is that a case of Goodhart's Law or is that just a case of, Facebook's goals are just not aligned with society's goals?

David Manheim: So there are a couple of things that are going on there. I saw a comment recently by Robin Hanson saying he doesn't understand why companies don't use AB tests or run experiments more often, because this is effective, why don't we do this?

And my immediate thought was, part of the reason why is because we don't have great metrics to run them with. So Facebook and other tech companies do. They are constantly running very sophisticated experiments internally to figure out what drives engagement most. Where “engagement” is measured by a metric that they've chosen, which isn't even necessarily what they want.

So Facebook may be incentivized to maximize number of users that engage with the platform every day because it's a metric that they report to investors. So it looks good. But part of what's happening is that they're confused about what it is that their users want.

So the studies seem to show that using social media makes people less happy. I don't think that that's something Facebook wants. I don't think that there's a conflict between what Facebook wants and what its users want there. I think there's a conflict between the easy to measure metrics that Facebook can look at, and what it is that it's optimizing for.

And the actual goals of both Facebook and the people using it, to provide a useful service that people are interested in, so that they'll look at it a bunch, and click on ads and make Facebook money. Or use Facebook a lot, so that Facebook can harvest their data, so that Facebook can sell it.

But whatever the business model is, I don't think that it's actually being served necessarily by the fact that the incentives are being misaligned.

Fake news is a really important example right now. Because it was absolutely inadvertent on the part of Facebook that their algorithms motivated people to share news that created filter bubbles, that led to people spreading fake news, that let foreign governments promote things. It was never intentional on Facebook's part to create that dynamic. The dynamic that existed was exploited by others.

So it's in some ways more difficult. Because this gets into a very complex multi-agent scenario, where the metric that Facebook is using is being gamed by governments and corporations that can figure out how it is that they can use that to manipulate the users of Facebook, whose goals and incentives are a third set of things that we care about.

Julia Galef: Right, that's a very complex one.

David Manheim: So it's a very complex case. I think that there are at least two or three different places where Goodhart phenomena are happening there. So it's a really good example, but it's a hard one to pull apart.

Julia Galef: Okay. All right. Let me give you one more. What about cases where governments pass regulations, like companies have to provide health care for employees who work at least 40 hours. So they set a discrete threshold... and then as a result, companies just have all their employees work 39 and a half hours.

Or they do some kind of complex... No, I'm going to stick to that example. They respond to discrete thresholds by getting as close to

the threshold as possible, and then not going over it. Does that count?

David Manheim: Yeah. So I actually was involved in a conversation, and dubbed this “Shorrock's Law of Limits,” based on something that a guy named Steven Shorrock said on Twitter, which is: If you put a limit on a measure, if the measure relates to efficiency, the limit gets used as a target.

So what happened here was really specifically that they said, "Look, here's the limit. If you hit 40, then you have to pay all this extra money." So what people did was they started, instead of saying like, "Oh well we'll have some people who we employ that we have for 20 hours and some that we have for 40," they started saying, "Great, let's get everybody at 38 hours or 39.8 hours so that we don't have to pay this because it saves us a ton of money to cut the 15 minutes off of our 40 a week worker. So of course we're going to do this."

And it's definitely closely related to Goodhart's Law. There's definitely a metric that's being looked at. But I don't know that... Kind of the central dynamic for Goodhart's Law is one where the metric stops being useful because of the way that it's being played with. I don't think that the regulators were trying to measure something specific with the “full time workers get health care, non full time workers don't get health care.” I think that they were using that as a convenient line.

So I'm not sure how helpful it is to start talking about which thing specifically qualifies as what we're calling a Goodhart effect or not. But there were definitely some dynamics in there that relate to the metric that's being used.

Julia Galef: Yeah. I mean, the reason this is helpful, or the reason I hope it's helpful for people, the reason it's helpful for me, is... looking at examples like this and whether you would call them Goodhart's Law helps highlight that there are multiple important phenomena going on. Where for example, one is lost purposes and confusion over what a metric should be. And a different phenomenon is, adversarial game theory where people have different incentives and will respond by serving their own goals instead of the spirit of the law. Stuff like that.

All right, so let me now try to summarize the different mechanisms that we've talked about:

One mechanism is this adversarial dynamic, which isn't really central to Goodhart's Law, like the nail factory in the Soviet Union.

Two is the vague goals and under-specified goals, making it difficult to figure out what metrics to set in an organization, for example.

Three is coordination difficulties where maybe you have... Like, it's clear what goal you're trying to pursue, as the head of the company, but it's really hard to come up with metrics that, when you implement them, will cause all the different departments to be optimizing for the right thing in a way that coordinates effectively.

And then... okay, so a fourth thing that I don't think we've talked about, that I want to ask you about, is genuine psychological confusion. Not over what my goal should be, but just getting... Like, coming to think that you should optimize for something that isn't what you wanted in the first place.

So for example, I've had some conversations with people about scientific progress and “Is scientific progress slowing down? How can we tell?”

And something that keeps happening in these conversations is that other people point to a metric of scientific progress that seems completely wrong to me. They point to the number of papers published. So they'll say, like, "Scientific progress is actually speeding up. Look the number of papers that are being published over the years -- or per researcher, even.” Researchers are publishing more papers per year over their careers than they used to.

And to me, that's so completely backwards. “Number of papers published” is an input, not an output. What we actually care about is the number of discoveries, or the number of important discoveries. And if number of papers published is going up, but the number of important discoveries is *not*, then that's even worse!

And so, I don't know what they would say to this accusation, but I feel like they've just gotten confused about what we care about, with respect to science.

David Manheim: So what I called the confusion that happens is people “reify” goals.

And reify is a term from psychology, where what happens is they take something that they think they see, that they think looks one way ,and they turn that into the thing itself. So you start with, “Oh, well we're trying to do science. Well, what is science? Science is comprised of people publishing papers. So papers are science, more papers are therefore more science.”

And it's not... I don't even think that that's wrong. More papers *are* more science. It's just that our goal isn't more science. Our goal is *advancing* science. We want progress. We don't just want things to happen.

So partially this is an example where I don't think that people are clear enough about what it is that your goals are. As a scientist, you're supposed to -- you know, even if you were to say, I think correctly, that science is about formalizing insights into the nature of reality so that you have better predictive models...

There's still a difference between better predictive models of the way in which sodium and oxygen chemically interact, and saying, "We have better models of how it is that bubbles form in water," versus better insight into how it is that when kids blow into a straw, it makes different things happen in the cup.

And you can publish a paper on any of the three of those. And I'm betting that the third one would get more media attention than the first two.

And that's a metric. I don't think it's the most useful one. But what you end up with, then, is this situation where people optimize for the easiest insights to find, the ones that are the best for their career. The ones that are going to help their citations the most. And all of those things are things that matter locally, because of the dynamics of the larger system, that aren't science.

But right back to your point, yes, people get fundamentally confused about what it is that their goal is. I had an example of this that I mentioned in one of my essays... which was, I noticed that I use Twitter a lot. It's a great thing to do when I'm trying to run an analysis and it's going to take 12 or 15 minutes to run and it's not enough time to do anything else. So I tweet, and I reply to people. And my behavior is naturally drawn to doing the things that are incentivized by Twitter.

So what I realized at one point was that if I put screenshots of the things that I link to in my tweets, they get a lot more engagement. And Twitter tells you, "This is how much engagement you have." And I'm like, "Oh, that's great. I definitely want more engagement."

And so I started doing that more. And at a certain point, I was looking at some of the metrics a little bit more, and I realized... yeah, it does drive more engagement. People click on the image, and they read the excerpt that I put -- and then don't click on the link.

So what I've done is cannibalized some of the actual people reading the thing that I think is important for them to read, into them reading the four lines that I thought were most interesting or most attention grabbing.

That's not a good trade off. That's not what I wanted at all. But I hadn't thought through this enough, and I had just grabbed the thing that I thought was useful, without a ton of reflection at all.

And that's exactly, in my mind, “Look, I reified engagement.” I want engagement. If I'm on Twitter and talking to people, I'd like them to interact with me. But the type of interaction that I want isn't that.

The same thing is true about clever, snarky comments on Twitter get lots of retweets, and lots of likes -- and probably drive away the type of person who I'd actually like to interact with on Twitter. Because it's not substantive and not interesting.

So if you're not really careful about what you're doing, then you absolutely end up with not your actual goals as what it is that you spend your time doing.

Julia Galef: Those don't even sound like reification mistakes to me. Those sound like you weren't paying close enough attention, or thinking carefully enough about what you wanted to optimize for. But I definitely, I have noticed reification mistakes in myself. Where for example, if I'm on a diet, what I actually care about is losing the weight. But it starts to feel like what I care about is making the number on the scale go down.

Those are obviously very closely linked -- but they're not exactly linked. And so what I will sometimes find myself tempted to do is to weigh myself when -- I don't drink a lot of water. Because I don't want the number on the scale to be higher because I've just drank a gallon of water. But that's not... The water isn't making me gain weight.

David Manheim: ... right, or after you use the bathroom. And then jump on the scale. Because maybe I've lost like a quarter pound, you're like, "That's the weight loss I'm looking for."

Julia Galef: And I can see myself doing it. But yeah.

David Manheim: So I actually, I think that in my case for Twitter, I had reified it in exactly that way. It wasn't just me not thinking about it. I think that part of it is, it's hard for somebody to see the difference between you not thinking...

So somebody sees you jumping on the scale, or not drinking very much water, and they think, "Oh she hasn't thought about her goal of what weight loss actually means very much. She just thinks that means that the number on the scale goes down." And that's not what happened. What happened is you do know exactly what it is your goal is, but your brain slipped a little bit. Because it's hard to pay attention to everything that your brain is doing on a low level.

Julia Galef: It's hard being a human.

David Manheim: Yeah, so I feel like that's exactly the right time to say, "Yeah, it's going to be harder to be an AI." All of these issues -- I just want to throw this in because I think it's key -- all of the issues that we have with Goodhart's Law, one of the key things that we can do to get around them is rely on judgment. Ask teachers to use their judgment about what it is they should teach a little bit more and follow guidelines a little bit less. Or ask people to just think a little bit about what it is they're trying to do, when you're giving them assignments at work. Just tell them, "Oh, by the way, you should push back if you think this is wrong."

Those are all the things that when you're automating systems you can't do. You can't tell Facebook's A/B test to, "By the way, think really quick about whether this is actually what it is we want."And so we end up in a much worse situation when we don't have all of the fuzzy stuff in people's head to fall back on.

Julia Galef: Yeah. The point being that even advanced artificial intelligences can't... They might develop something we would call judgment, but it isn't going to be a close match for the judgments that we would make, as humans. Because there's a lot of implicit stuff in our quote-unquote “utility functions,” as humans, that we can't easily transfer to an AI.

David Manheim: Yeah. There are a couple of different dynamics that apply here and this is one of those places where I've had a bunch of discussions on Less Wrong and other places, with people who are focused much more specifically on AI.

And I don't think that there's a lot of clarity on exactly where to draw the line between AI not being aligned, and AI gaming targets, and AI falling prey to Goodhart's Law. And I'm not sure that the lines between these are clear to... They're certainly not clear to me. I don't think that they're clear to a lot of the people who are more actively working on this. If they seem clear to any of the listeners, it'd be great for them to write a blog post.

Julia Galef: Yeah, yeah. No, that's a good point. Those tend to get... They're not clear to me either. I'll say that.

Would you say the difficulty of empowering people to just use their judgment is part of why startups often struggle when they scale up, to become larger, more established companies? Because coordination is so much harder, and you can't just tell people to use their judgment?

David Manheim: There are a couple of things that go on there. That's part of it. I mean, that's a complicated topic that I've thought a lot about.

Some of the things that go wrong are simply that when things get bigger, it's not that it's harder to tell people to use their judgment. It's that it reflects worse on the people in charge when they don't go well, and the people not in charge have more difficulty doing the things.

So, if you're in a startup, if there are three people in the room, then the CEO tells somebody, "Oh, by the way, this is what we should do," and the junior person in the room -- who's the second guy in the company -- says, "Maybe I should take this riskier thing and do this."

There's a great study that somebody did a long time ago in management. Where they asked a bunch of senior managers -- not the CEO, but senior managers – “If you had a choice between these two projects, where one of them has a 50% chance of failing and a 50% chance of quadrupling the money that you invest; Or a project that has a 99.9% chance of returning 12% on your money and a .1% chance of returning only 1% on the money, which one do you do?”

And all of the senior managers go, "That second one sounds great."

Julia Galef: Really?

David Manheim: Because what happens if you invest 50% of your annual budget in the project and it quadruples? You get a nice bonus, you get recognition, everybody's happy with you. And what if it fails? You probably get fired.

Julia Galef: Right, right.

David Manheim: So, they ask the CEO, "Which one do you want people to do?" And the CEO says, "What do you mean? Of course, I want them to do the first one. This is a crazy question. Why would any of my subordinates not do the first thing?"

And then they said, "And so if you found out that one of your subordinates did that first thing and it failed, what would you do?"

And the CEO is like, "Oh, they'd get fired."

Julia Galef: Oh, man. Well, I guess we've identified the fault in the system. This is not a case where you look at a complex system and you're like, "Where are things going wrong? I can't find the broken parts..."

David Manheim: The problem here is... just to narrow it down a little bit, the problem here is that when you're in a large company, people can't be informal about things. They have to have fully delegated the responsibility. And the person who's making these decisions has to fully take responsibility for the outcome.

And so you end up in this situation where you haven't actually aligned incentives with what it is you want to. And most of the time the reason why is because actually spending the time to really align incentives would be much more work than it's worth. But you definitely get misalignments because of that.

So, as companies get bigger, some of what happens is some of the junior people, who are really used to being able to say, "I'm going to take this really risky move because I know the CEO will have my back" ...and they do it. And the CEO is like, "I would have their back, but now we have investors, and they're screaming that somebody needs to be fired. So, I don't know what I'm supposed to do here... but the guy's going to end up being fired."

So, you end up with just different dynamics because of the fact that it's changed. And some of those have a little bit to do with the metrics and how you align people. And some of them have to do with other factors about how it is that organizations work.

Julia Galef: When I was reading one of your posts about how metrics kind of stand in for people using their judgment and intuition, especially in large complex systems where you're responsible for one piece of it... I had this idea for how you could get around that problem in a large company.

Which I'm sure is wrong. Because the chances I would have come up with a plan for organizational theory that people aren't already doing are low. But why don't you tell me why this plan wouldn't work?

So, to be clear, the problem that I'm trying to solve is the CEO, let's stipulate, has in their mind a complete understanding of what they

would like everyone in the company to do. If they could just look at each person's work-

David Manheim: If they could do all the jobs for them.

Julia Galef: Right. Exactly. They have in their mind what we as a company are trying to optimize for. But it's hard to specify it in a clear way, a simple way, such that they can just give everyone a task and have everyone go off and do the thing, and now the company will just be optimizing. That's just too hard.

So, they try to do something kind of like this. They give people metrics, and those managers give their employees metrics and so on so forth. But it's just so crude that you end up having lost purposes. Exactly. That's the problem I'm trying to solve.

So, let's say there's the CEO, and then below them a tier of upper managers, then middle managers, then lower managers, and then more employees.

Can you not have any metrics at all, but instead you just have the CEO "check the work" of the upper managers below them? By which I mean, say the CEO looks at maybe a sample of 10% of the decisions that each upper manager makes. And because it's only 10%, the CEO can spend the time to understand that decision.

Say the decision is the upper manager is evaluating the middle managers below them. And the CEO looks at 10% of those decisions and says, "Here's how I would have made that decision, given my perfect model in my head of what we're optimizing for. And I'm going to reward or punish you based on how close yours was, to what I would've done."

And then in turn, the upper managers do the same with the middle managers below them. They scrutinize 10% of the decisions that each middle manager makes, about how to evaluate the lower managers. And so on and so forth. So, it's basically like reinforcement learning. It's propagating the CEO's mental model --

David Manheim: Right. We should use what works for AI more in the line of people.

Julia Galef: In companies. Yeah, that's what I'm... yeah.

David Manheim: And I think that some of the intuition there is reasonable. I'll point out a couple of reasons that, in practice, this is a really, really problematic thing to do.

The first one is when we do this with reinforcement learners, there's some idea about what the goal is at the beginning. So, if the CEO gives a bunch of speeches and say, "Look, this is what we want to do. Everybody pay attention to the speech, and then figure out what it is that I want. And I'll look at some of your work and check and see if it's actually what I want you to do, and I'll make judgements based on that. And some of you may be fired and some of you may get big bonuses, because you're going to be more or less aligned with what it is I think that needs to happen"

... people have a hard time operationalizing that. It's the kind of thing that if you were working in a company like that, as a mid-level manager, you would constantly be terrified that you're doing something wrong. But you don't know what because you haven't been given a clear goal. You haven't been given a clear metric. So, you're in this really weird situation --

Julia Galef: But can't you be reassured by the fact that you're not going to get fired? The rewards and punishments are continuous, they're not discrete.

David Manheim: So, optimize less on the thing, so then you just have less pressure on them to do the right thing. So, yes, there are some things that are like that.

The other part of this is, if you look at what companies did before the era of scientific management, which is going back to like a century ago, maybe a little bit more... before they had the idea of having metrics, and before they actually measured things much, this is kind of what happened.

And most companies kind of... Most of what happened ended up being judged on biases, that weren't actually how good a job you did. But the CEO, instead of saying, "Oh, this is exactly what I would have done" ends up saying -- because he doesn't have any concrete yardstick -- "This guy seems likable..."

Julia Galef: Right. Or he flatters me or... Yeah.

David Manheim: "... And I don't have any really clear reason to say that he did something wrong. And this guy, when we were golfing the other week, kept on slicing the ball into the lake, and it was really annoying. So, I'm not thinking about that, but I kind of don't like this guy. So, anything that he did is kind of bad anyway."

So, if you don't have clear metrics, then you do end up with people's biases taking a huge role. And it's not necessarily true that the

biases that people have would overwhelm their actual judgment, but...

And this is the last point about why it is that I think that this is a certainly really bad idea in practice. Which is that, I would guess... I'm not a lawyer. I would guess that the lawsuits about somebody not getting the bonus, or getting fired, or anything like that, in a system like this, would be impossible for the lawyers to defend.

And everybody ends up furious, and you'd end up losing tons of money because you spent 50% of the profit of the company defending against the four lawsuits from the people who actually probably should have been fired, but because you don't have a defensible system for explaining why...

: So, there are elements there that I think would be useful, but I also think that... Yeah, this is not something I would tell a CEO to do.

Julia Galef: Okay, fine. So, maybe I shouldn't run society just yet. But if I think about it a little longer, maybe I'll be able to centrally plan how organizations work!

David Manheim: I don't think that it's essentially a bad idea. I think that some amount of doing exactly that is what good managers do. What good managers do is they stop by.

I've heard this specifically about Elon Musk that he stops by people's desks. I had a friend who was at SpaceX and he'll stop by. You're not paying attention and he'll lean over your shoulder and be like, "So, why do we shape it like that?" And you'll look over your shoulder and be like, oh, the --

Julia Galef: “Oh, hello, Elon.” Yeah.

David Manheim: Yeah. And you'll like walk through your thinking, and about 90% of the time, he'll be like, "Oh, that's good." And about 10% of the time, he'll be like, "Wait, no. We should be able to do this, this and this," and he'll want you to defend your...

It's not like he's like, "No, fix it. You did it wrong," but he'll want you to explain why it is that you didn't do that. And he's a really bright guy, so he's not asking dumb questions.

And that actually helps. I think that that's a very extreme example, but good managers do stop by and say, "Hey, so I was looking at the work you handed in and this seems a little bit off" or "This seems

really great. You did a good job. Keep on going." That is what they're supposed to be doing to some extent.

But that's on top of the metrics. That's not... Yeah.

Julia Galef: Yeah. So, we'll wrap up in a minute, but are there any effective ways to get around Goodhart's law that we haven't talked about?

David Manheim: Yeah. Actually, I have a paper about this recently. I'm trying to figure out where to submit it... but basically, there were a couple of really specific strategies.

One is, make a bunch of metrics instead of just one. And figure out if you can... Hopefully they fail in different ways. So that if you look at all of them, you don't end up messing up as badly as when you incentivize people to build lots of little nails

Julia Galef: Are you just trading off... You're making it more robust to the kind of problems that Goodhart's law causes, but in exchange, you're building in more of a role for your own intuition/biases? Because you have to decide how to weigh those different metrics against each other?

David Manheim: So, you could even specify how you weigh them beforehand. That's not a problem. It turns out it's really hard to game complex metrics compared to gaming simple ones. Which is not to say that people will not spend some effort doing it, but it's more complex.

That's a benefit and a problem. Because you don't want your goals to be so complex that people can't figure out how to accomplish them -- but you do want them to be complex enough that the easiest way to do them is to actually do the thing you're supposed to do.

Julia Galef: Right. Yes.

David Manheim: So, that's the trade off.

The next piece, I think, that's really important on how to deal with this, is don't put too much optimization pressure on things.

Finance does this horribly, where they will almost explicitly say, "By the way, your bonus is going to be about 10% of the profits you pull in in a given year."

Julia Galef: Yeah.

David Manheim: Well, it's really clear what it is you're optimizing for, and it's short-term profits over the course of a year. So, go forth and take risks you shouldn't.

If you push really hard to optimize on a goal... So, if you give people $50 gift certificates, or a Visa gift card when they do the thing that you want them to, that may be too little optimization pressure, but you're probably not going to fall prey to Goodhart's law to any really significant extent.

If you put five times their annual salary riding on the metric, then yeah, you're going to end up messing things up.

Julia Galef: Right.

David Manheim: The next thing is, yeah, relying on people's judgment isn't a bad idea. And there's a book, The Tyranny of Metrics, that basically spent 200 pages saying, "So, we should rely on people's judgment more."

... That's not fair, because places where people use metrics, it improves things. The world has gotten a lot better now that we have people actually measuring the results of what they do.

There are some downsides if you push too hard in places where you're not 100% sure what it is you're doing -- so there are good reasons to be careful, but... don't abandon metrics, but sometimes abandon metrics. This is not a good place to use metrics here. There are places where that's going to be true.

So, I think that those are the big ones that I would say people should be paying attention to. And a lot of this is... If I had hard and fast rules for where you should and shouldn't use metrics, I'd be thrilled, but it's not quite that simple.

Julia Galef: I would be shocked and impressed if we lived in a world where there were such hard and fast rules. And probably once those hard and fast rules became well known, they would be gamed and so they would no longer be applicable. So... Great. Well, David, before we wrap up, I mentioned that I was curious if there was a particular book or other resource that you could point to that was particularly influential on your thinking or your life.

David Manheim: So, I'm going to skip the easy examples that I have and not talk about Peter Singer or the Sequences or anything, and go with... I think I mentioned it earlier, but *Bureaucracy* by James Q. Wilson.

Julia Galef: Oh, yeah. Can you talk a little more about that?

David Manheim: The subtitle is “What Government Agencies Do And Why They Do It,” and it's really very readable. I mean, it's a little bit academic, but it's really very readable. And it actually goes through, “Hey, this is why some government agencies do a fantastic job.”

Social security administration is great. They have a very clearly defined job -- they send out checks, the checks get there on time. Everybody knows what they're supposed to do, and it gets done, and it's fantastic.

And there are some government agencies where it's really hard, and there's horrible bureaucracy, and nobody knows how to fix it. And there are good reasons why.

Julia Galef: Do you recall any examples in that category?

David Manheim: So, the first thing I would say is if you look at, for instance, the US military... the primary reason that it's not efficient, and people complain about the fact that there are all sorts of things that are not efficient, is because it's about as efficient as you would expect for an organization that's 10 times the size of the largest company in the world. It's huge.

Julia Galef: Right, Yeah. That's a good point.

David Manheim: There's no way to manage that.

And what is their output? And peacetime militaries... He talks about this a lot. Peacetime militaries are in a really bad situation, where what are they supposed to be doing? Getting ready to do a good job at something in the future, in an undefined future scenario. Well, how do you measure that? How do you figure out what they're supposed to be doing?

So, it's a really hard situation to be in, and there are ways to do it slightly better and slightly worse, but there are good reasons to say, "Yeah, so that's why it's hard to figure out what it is that this bureaucracy should be doing."

So, it has a lot in there about kind of better understanding what it is that happens, specifically in government. And I think that it's really useful. Because people like to dump on government for being inefficient... and I think that they're right in a lot of places, but there are good reasons why it works the way it does.

But it's also really valuable, and people use it a lot in business schools to talk about how businesses end up in some of the same places. So, it's a really... Highly recommend it.

Julia Galef: Excellent. And would you say that in your trajectory in particular, was it mostly influential in getting you interested in analyzing organizations and systems through these lenses?

Or was it like, you used to view government as just incompetent, and the book caused you to recognize some of the hard problems that government is trying to solve, that you didn't see?

David Manheim: I was in grad school learning a lot about this, so it definitely wasn't as simple as “I used to think...” But there were a lot of places where I updated really significantly about where the problems were, and what types of things you need to think about to understand them better.

And there are some tools in there that really do help you, like “Oh, this is why this is hard,” or “These are the types of thing that people have tried that don't work, or that do work.”

Julia Galef: Nice. I really appreciate books where I come away with kind of a tool for analyzing things. Or general questions to ask myself in trying to understand other completely unrelated things. Or seemingly unrelated things, to the topic of the book. That's a treat.

David Manheim: So, if you want to understand bureaucracies, it's really highly recommended. It won't help you get the phone company to transfer your number faster or whatever, but it will help you understand why it is that it's so hard.

Julia Galef: I wonder if there are any disgruntled Amazon reviewers who are like, "I was hoping this would help me figure out how to deal with Comcast bureaucracy. One star."

David Manheim: How to get Amazon to refund me.

Julia Galef: Right, exactly. I encourage our listeners to follow David on Twitter for more scintillating insights like those you've just heard. His Twitter handle is David Manheim. That's D-A-V-I-D M-A-N-H-E-I-M. David Manheim.

All right, well, David, thank you so much for coming on the show. It's been an enlightening hour. I appreciate it.

David Manheim: Thank you.

Julia Galef: This concludes another episode of Rationally Speaking. Join us next time for more explorations on the borderlands between reason and nonsense.