CHAPTER 22

Open Data and Algorithmic Regulation

By Tim O'Reilly

Regulation is the bugaboo of today's politics. We have too much of it in most areas, we have too little of it in others, but mostly, we just have the wrong kind, a mountain of paper rules, inefficient processes, and little ability to adjust the rules or the processes when we discover the inevitable unintended results.

Consider, for a moment, regulation in a broader context. Your car's electronics regulate the fuel-air mix in the engine to find an optimal balance of fuel efficiency and minimal emissions. An airplane's autopilot regulates the countless factors required to keep that plane aloft and heading in the right direction. Credit card companies monitor and regulate charges to detect fraud and keep you under your credit limit. Doctors regulate the dosage of the medicine they give us, sometimes loosely, sometimes with exquisite care, as with the chemotherapy required to kill cancer cells while keeping normal cells alive, or with the anesthesia that keeps us unconscious during surgery while keeping vital processes going. ISPs and corporate mail systems regulate the mail that reaches us, filtering out spam and malware to the best of their ability. Search engines regulate the results and advertisements they serve up to us, doing their best to give us more of what we want to see.

What do all these forms of regulation have in common?

1. A deep understanding of the desired outcome

2. Real-time measurement to determine if that outcome is being achieved

3. Algorithms (i.e. a set of rules) that make adjustments based on new data

4. Periodic, deeper analysis of whether the algorithms themselves are correct and performing as expected.

There are a few cases—all too few—in which governments and quasi-governmental agencies regulate using processes similar to those outlined above. Probably the best example is the way that central banks regulate the money supply in an attempt to manage interest rates, inflation, and the overall state of the economy. Surprisingly, while individual groups might prefer the US Federal Reserve to tighten or loosen the money supply at a different time or rate than they do, most accept the need for this kind of regulation.

Why is this?

1. The desired outcomes are clear

2. There is regular measurement and reporting as to whether those outcomes are being achieved, based on data that is made public to everyone

3. Adjustments are made when the desired outcomes are not being achieved

Contrast this with the normal regulatory model, which focuses on the rules rather than the outcomes. How often have we faced rules that simply no longer make sense? How often do we see evidence that the rules are actually achieving the desired outcome?

Sometimes the "rules" aren't really even rules. Gordon Bruce, the former CIO of the city of Honolulu, explained to me that when he entered government from the private sector and tried to make changes, he was told, "That's against the law." His reply was "OK. Show me the law." "Well, it isn't really a law. It's a regulation." "OK. Show me the regulation." "Well, it isn't really a regulation. It's a policy that was put in place by Mr. Somebody twenty years ago." "Great. We can change that!"

But often, there really is a law or a regulation that has outlived its day, an artifact of a system that takes too long to change. The Obama Administration has made some efforts to address this, with a process of

both "regulatory lookback" to eliminate unnecessary regulations, and an increased effort to quantify the effect of regulations (White House, 2012).

But even this kind of regulatory reform doesn't go far enough. The laws of the United States have grown mind-bogglingly complex. The recent healthcare reform bill was nearly two thousand pages. The US Constitution, including two hundred years worth of amendments, is about twenty-one pages. The National Highway Bill of 1956, which led to the creation of the US Interstate Highway system, the largest public works project in history, was twenty-nine pages.

Laws should specify goals, rights, outcomes, authorities, and limits. If specified broadly, those laws can stand the test of time.

Regulations, which specify how to execute those laws in much more detail, should be regarded in much the same way that programmers regard their code and algorithms, that is, as a constantly updated toolset to achieve the outcomes specified in the laws.

Increasingly, in today's world, this kind of algorithmic regulation is more than a metaphor. Consider financial markets. New financial instruments are invented every day and implemented by algorithms that trade at electronic speed. How can these instruments be regulated except by programs and algorithms that track and manage them in their native element in much the same way that Google's search quality algorithms, Google's "regulations", manage the constant attempts of spammers and black hat SEO experts to game the system?

Revelation after revelation of bad behavior by big banks demonstrates that periodic bouts of enforcement aren't sufficient. Systemic malfeasance needs systemic regulation. It's time for government to enter the age of big data. Algorithmic regulation is an idea whose time has come.

Open Data and Government as a Platform

There are those who say that government should just stay out of regulating many areas, and let "the market" sort things out. But there are

many ways in which bad actors take advantage of a vacuum in the absence of proactive management. Just as companies like Google, Microsoft, Apple, and Amazon build regulatory mechanisms to manage their platforms, government exists as a platform to ensure the success of our society, and that platform needs to be well regulated!

Right now, it is clear that agencies like the SEC just can't keep up. In the wake of Ponzi schemes like those of Bernie Madoff and Allen Stanford, the SEC has now instituted algorithmic models that flag for investigation hedge funds whose results meaningfully outperform those of peers using the same stated investment methods. But once flagged, enforcement still goes into a long loop of investigation and negotiation, with problems dealt with on a case-by-case basis. By contrast, when Google discovers via algorithmic means that a new kind of spam is damaging search results, they quickly change the rules to limit the effect of those bad actors. We need to find more ways to make the consequences of bad action systemic, rather than subject to haphazard enforcement.

This is only possible when laws and regulations focus on desired outcomes rather than the processes used to achieve them.

There's another point that's worth making about SEC regulations. Financial regulation depends on disclosure - data required by the regulators to be published by financial firms in a format that makes it easy to analyze. This data is not just used by the regulators themselves, but is used by the private sector in making its own assessments of the financial health of firms, their prospects, and other financial decisions. You can see how the role of regulators in requiring what is, in effect, open data, makes the market more transparent and self-policing.

You can also see here that the modernization of how data is reported to both the government and the market is an important way of improving regulatory outcomes. Data needs to be timely, machine readable, and complete. (See Open Government Working Group, 2007.) When reporting is on paper or in opaque digital forms like PDF, or released only quarterly, it is much less useful.

When data is provided in re-usable digital formats, the private sector

can aid in ferreting out problems as well as building new services that provide consumer and citizen value. This is a goal of the US Treasury Department's "Smart Disclosure" initiative (see http://www.data.gov/consumer/page/consumer-about). It is also central to the efforts of the new Consumer Financial Protection Bureau.

When government regulators focus on requiring disclosure, that lets private companies build services for consumers, and frees up more enforcement time to go after truly serious malefactors.

Regulation Meets Reputation

It is true that "that government governs best that governs least." But the secret to "governing least" is to identify key outcomes that we care about as a society—safety, health, fairness, opportunity—encode those outcomes into our laws, and then create a constantly evolving set of regulatory mechanisms that keep us on course towards them.

We are at a unique time when new technologies make it possible to reduce the amount of regulation while actually increasing the amount of oversight and production of desirable outcomes.

Consider taxi regulation. Ostensibly, taxis are regulated to protect the quality and safety of the consumer experience, as well as to ensure that there are an optimal number of vehicles providing service at the time they are needed. In practice, most of us know that these regulations do a poor job of ensuring quality or availability. New services like Uber and Hailo work with existing licensed drivers, but increase their availability even in less-frequented locations, by using geolocation on smartphones to bring passengers and drivers together. But equally important in a regulatory context is the way these services ask every passenger to rate their driver (and drivers to rate their passenger). Drivers who provide poor service are eliminated. As users of these services can attest, reputation does a better job of ensuring a superb customer experience than any amount of government regulation.

Peer-to-peer car services like RelayRides, Lyft, and Sidecar go even further, bypassing regulated livery vehicles and allowing consumers to provide rides to each other. Here, reputation entirely replaces regu-

lation, seemingly with no ill effect. Governments should be studying these models, not fighting them, and adopting them where there are no demonstrable ill effects.

Services like AirBnB provide similar reputation systems that protect consumers while creating availability of lodging in neighborhoods that are often poorly served by licensed establishments.

Reputation systems are a great example of how open data can help improve outcomes for citizens with less effort by overworked regulators and enforcement officials.

Sites like Yelp provide extensive consumer reviews of restaurants; those that provide poor food or service are flagged by unhappy customers, while those that excel are praised.

There are a number of interesting new projects that attempt to combine the reach and user-friendliness of consumer reputation systems with government data. One recent initiative, the LIVES standard, developed by San Francisco, Code for America, and Yelp, brings health department inspection data to Yelp and other consumer restaurant applications, using open data to provide even more information to consumers. The House Facts standard does the same with housing inspection data, integrating it with internet services like Trulia

Another interesting project that actually harnesses citizen help (rather than just citizen opinion) by connecting a consumer-facing app to government data is the PulsePoint project, originally started by the San Ramon, California fire department. After the fire chief had the dismaying experience of hearing an ambulance pull up to the restaurant next door to the one in which he was having lunch with staff including a number of EMR techs, he commissioned an app that would allow any citizen with EMR training to receive the same dispatch calls as officials.

The Role of Sensors in Algorithmic Regulation

Increasingly, our interactions with businesses, government, and the built environment are becoming digital, and thus amenable to creative

forms of measurement, and ultimately algorithmic regulation.

For example, with the rise of GPS (not to mention automatic speed cameras), it is easy to foresee a future where speeding motorists are no longer pulled over by police officers who happen to spot them, but instead automatically ticketed whenever they exceed the speed limit.

Most people today would consider that intrusive and alarming. But we can also imagine a future in which that speed limit is automatically adjusted based on the amount of traffic, weather conditions, and other subjective conditions that make a higher or lower speed more appropriate than the static limit that is posted today. The endgame might be a future of autonomous vehicles that are able to travel faster because they are connected in an invisible web, a traffic regulatory system that keeps us safer than today's speed limits. The goal, after all, is not to have cars go slower than they might otherwise, but to make our roads safe.

While such a future no doubt raises many issues and might be seen by many as an assault on privacy and other basic freedoms, early versions of that future are already in place in countries like Singapore and can be expected to spread more widely.

Congestion pricing on tolls, designed to reduce traffic to city centers, is another example. Systems such as those in London where your license plate is read and you are required to make a payment will be replaced by automatic billing. You can imagine the costs of tolls floating based not just on time of day but on actual traffic.

Smart parking meters have similar capabilities—parking can cost more at peak times, less off-peak. But perhaps more importantly, smart parking meters can report whether they are occupied or not, and eventually give guidance to drivers and car navigation systems, reducing the amount of time spent circling while aimlessly looking for a parking space.

As we move to a future with more electric vehicles, there are already proposals to replace gasoline taxes with miles driven—reported, of course, once again by GPS.

Moving further out into the future, you can imagine public transpor-

tation reinventing itself to look much like Uber. It's a small leap from connecting one passenger and one driver to picking up four or five passengers all heading for the same destination, or along the same route. Smartphone GPS sensors and smart routing algorithms could lead to a hybrid of taxi and bus service, bringing affordable, flexible public transportation to a much larger audience.

The First Step is Measurement

Data driven regulatory systems need not be as complex as those used by Google or credit card companies, or as those imagined above. Sometimes, it's as simple as doing the math on data that is already being collected and putting in place new business processes to act on it.

For example, after hearing of the cost of a small government job search engine for veterans ($5 million per year), I asked how many users the site had. I was told "A couple of hundred." I was understandably shocked, and wondered why this project was up for contract renewal. But when I asked a senior official at the General Services Administration if there were any routine process for calculating the cost per user of government websites, I was told, "That would be a good idea!" It shouldn't just be a good idea; it should be common practice!

Every commercial website not only measures its traffic, but constantly makes adjustments to remove features that are unused and to test new ones in their place. When a startup fails to gain traction with its intended customers, the venture capitalists who backed it either withdraw their funding, or "pivot" to a new approach, trying multiple options till they find one that works. The "lean startup" methodology now widely adopted in Silicon Valley considers a startup to be "a machine for learning," using data to constantly revise and tune its approach to the market. Government, by contrast, seems to inevitably double down on bad approaches, as if admitting failure is the cardinal sin.

Simple web metrics considered as part of a contract renewal are one simple kind of algorithmic regulation that could lead to a massive simplification of government websites and reduction of government IT costs. Other metrics that are commonly used on the commercial web

include time on site; abandon rate (people who leave without completing a transaction); and analysis of the paths people use to reach the desired information.

There is other data available as well. Many commercial sites use analysis of search queries to surface what people are looking for. The UK Government Digital Service used this technique in their effort to redesign the Gov.UK site around user needs rather than around the desires of the various cabinet offices and agencies to promote their activities. They looked what people were searching for, and redesigned the site to create new, shorter paths to the most frequently searched-for answers. (Code for America built a site for the city of Honolulu, Honolulu Answers, which took much the same approach, adding a citizen "write-a-thon" to write new, user friendly content to answer the most asked questions.)

This is a simpler, manual intervention that copies what Google does algorithmically when it takes search query data into account when evaluating which results to publish. For example, Google looks at what they call "long clicks" versus "short clicks." When the user clicks on a search result and doesn't come back, or comes back significantly later, indicating that they found the destination link useful, that is a long click. Contrast that to a short click, when users come back right away and try another link instead. Get enough short clicks, and your search result gets demoted.

There are many good examples of data collection, measurement, analysis, and decision-making taking hold in government. In New York City, data mining was used to identify correlations between illegal apartment conversions and increased risk of fires, leading to a unique cooperation between building and fire inspectors. In Louisville, KY, a department focused on performance analytics has transformed the culture of government to one of continuous process improvement.

It's important to understand that these manual interventions are only an essential first step. Once you understand that you have actionable data being systematically collected, and that your interventions based on that data are effective, it's time to begin automating those interventions.

There's a long way to go. We're just at the beginning of thinking about

how measurement, outcomes, and regulation come together.

Risks of Algorithmic Regulation

The use of algorithmic regulation increases the power of regulators, and in some cases, could lead to abuses, or to conditions that seem anathema to us in a free society. "Mission creep" is a real risk. Once data is collected for one purpose, it's easy to imagine new uses for it. We've already seen this in requests to the NSA for data on American citizens originally collected for purposes of fighting overseas terrorism being requested by other agencies to fight domestic crime, including copyright infringement! (See Lichtblau & Schmidt, 2013.)

The answer to this risk is not to avoid collecting the data, but to put stringent safeguards in place to limit its use beyond the original purpose. As we have seen, oversight and transparency are particularly difficult to enforce when national security is at stake and secrecy can be claimed to hide misuse. But the NSA is not the only one that needs to keep its methods hidden. Many details of Google's search algorithms are kept as a trade secret lest knowledge of how they work be used to game the system; the same is true for credit card fraud detection.

One key difference is that a search engine such as Google is based on open data (the content of the web), allowing for competition. If Google fails to provide good search results, for example because they are favoring results that lead to more advertising dollars, they risk losing market share to Bing. Users are also able to evaluate Google's search results for themselves.

Not only that, Google's search quality team relies on users themselves—tens of thousands of individuals who are given searches to perform, and asked whether they found what they were looking for. Enough "no" answers, and Google adjusts the algorithms.

Whenever possible, governments putting in place algorithmic regulations must put in place similar quality measurements, emphasizing not just compliance with the rules that have been codified so far but with the original, clearly-specified goal of the regulatory system. The data used to make determinations should be auditable, and whenever possi-

ble, open for public inspection.

There are also huge privacy risks involved in the collection of the data needed to build true algorithmic regulatory systems. Tracking our speed while driving also means tracking our location. But that location data need not be stored as long as we are driving within the speed limit, or it can be anonymized for use in traffic control systems.

Given the amount of data being collected by the private sector, it is clear that our current notions of privacy are changing. What we need is a strenuous discussion of the tradeoffs between data collection and the benefits we receive from its use.

This is no different in a government context.

In Conclusion

We are just at the beginning of a big data algorithmic revolution that will touch all elements of our society. Government needs to participate in this revolution.

As outlined in the introduction, a successful algorithmic regulation system has the following characteristics:

1. A deep understanding of the desired outcome

2. Real-time measurement to determine if that outcome is being achieved

3. Algorithms (i.e. a set of rules) that make adjustments based on new data

4. Periodic, deeper analysis of whether the algorithms themselves are correct and performing as expected.

Open data plays a key role in both steps 2 and 4. Open data, either provided by the government itself, or required by government of the private sector, is a key enabler of the measurement revolution. Open data also helps us to understand whether we are achieving our desired

objectives, and potentially allows for competition in better ways to achieve those objectives.

About the Author

Tim O'Reilly is the founder and CEO of O'Reilly Media Inc., thought by many to be the best computer book publisher in the world. O'Reilly Media also hosts conferences on technology topics, including the O'Reilly Open Source Convention, Strata: The Business of Data, and many others. Tim's blog, the O'Reilly Radar "watches the alpha geeks" to determine emerging technology trends, and serves as a platform for advocacy about issues of importance to the technical community. Tim is also a partner at O'Reilly AlphaTech Ventures, O'Reilly's early stage venture firm, and is on the board of Safari Books Online, PeerJ, Code for America, and Maker Media, which was recently spun out from O'Reilly Media. Maker Media's Maker Faire has been compared to the West Coast Computer Faire, which launched the personal computer revolution.

References

Lichtblau, E., & Schmidt, M.S. (2013, August 3). Other Agencies Clamor for Data N.S.A. Compiles. The *New York Times*. Retrieved from http://www.nytimes.com/2013/08/04/us/other-agencies-clamor-for-data-nsa-compiles.html

Open Government Working Group. (2007, December 8). 8 Principles of Open Government Data. Retrieved from http://www.opengovdata.org/home/8principles

The White House. (2012). As Prepared for Delivery: Regulation: Looking Backward, Looking Forward - Cass R. Sunstein. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/inforeg/speeches/regulation-looking-backward-looking-forward-05102012.pdf