All Systems will be Gamed: Exploitive Behavior in Economic and Social Systems

W. Brian Arthur

SFI WORKING PAPER: 2014-06-016

SANTA FE INSTITUTE

All Systems will be Gamed:
Exploitive Behavior in Economic and Social Systems

W. Brian Arthur

Abstract

After the 2008 Wall Street crash, it became clear to economists that financial systems, along with other social and economic systems, are not immune to being manipulated by small groups of players to their own advantage. Given this, two natural questions arise. For a particular policy design or proposed economic system, can such manipulation be foreseen in advance and possibly prevented? And can we design methods—possibly automatic ones—that would test proposed policy systems for possible failure modes and for their vulnerability to possible manipulation, and thereby prevent such behavior in the future?

The paper argues that exploitive behavior within the economy is by no means rare and falls into specific classes; that policy studies can be readily extended to investigate the possibility of the policy's being "gamed"; and that economics needs a strong sub-discipline of failure-mode analysis, parallel to the successful failure-mode-analysis disciplines within structural engineering and aircraft design.

Key Words Economic Policy; financial crises; manipulation; exploitation; failure-mode analysis; nonequilibrium economics

To appear in *Complexity and the Economy*,
W. B Arthur, Oxford University Press, 2014

All Systems will be Gamed:
Exploitive Behavior in Economic and Social Systems

W. Brian Arthur [1]

There is a general rule in social and economic life: Given any system, people will find a way to exploit it. Or to say this more succinctly: All systems will be gamed. This is not a universal rule and it is certainly not a physical law; it is merely an observational truism. Given any governmental system, any legal system, regulatory system, financial system, election system, set of policies, set of organizational rules, set of international agreements, people will find unexpected ways to manipulate it to their advantage. "Show me a 50-foot wall," said Arizona's governor Janet Napolitano, speaking in 2005 of illegal immigration at the US-Mexico border, "and I'll show you a 51-foot ladder."

Foreseeing 51-foot ladders may not be particularly challenging—Napolitano is just making a wry political point. But anticipating more generally how exploitive behavior can arise in a given policy system *is* challenging; there are many ways in which systems can be exploited and some are by no means obvious. Yet we do need to foresee possible manipulations, not least because they can sometimes have disastrous consequences. Consider the aftermath of Russia's 1990 transition from planned socialism to capitalism, in which a small number of well-positioned players seized control of the state's newly freed assets. Or consider California's 2000 freeing of its energy market, in which a small number of suppliers were able to manipulate the market to the detriment of the state. Or consider Iceland's banking system in 2008, where a few financial players who had taken control of the state's banks used depositors' assets to speculate in overseas property markets and ran the banks into insolvency. Or consider Wall Street's loosely regulated mortgage-backed securities market in 2008, in which easy credit and complicated derivative products built a highly unstable structure that

[1] External Professor, Santa Fe Institute, and Visiting Researcher, Intelligent Systems Lab, PARC, Palo Alto. I thank Paul Maglio, Peter Haas, and Pat Selinger at IBM for useful comments; and also Sai Hung Cheung and Daria Rothmaier. An earlier version of this paper was written in 2010 when I was an IBM Faculty Fellow and member of IBM Almaden's Smarter Planet Platform for Analysis and Simulation of Health (SPLASH) group.

spectacularly collapsed. All these systems were manipulated—some were "gamed," to use a stronger term. All, in retrospect posed incentives that rendered them open to manipulation—and all careened into eventual system breakdowns.

This raises an obvious question. Given that economics is sophisticated and that economists study proposed policy systems in advance, how could these various economic disasters have happened? In the cases I mentioned some economists did indeed foresee possibilities for future exploitation and warn of these. But such warnings normally have little effect. The reason is that economics, in the way it is practiced, contains a bias that inhibits economists from seeing future potential exploitation. Economic analysis assumes equilibrium of the system in question, and by definition equilibrium is a condition where no agent has any incentive to diverge from its present behavior. It follows that for any system being studied invasive or exploitive behavior cannot happen: If a system could be invaded, some agents would be initiating new behavior, and the system could not have been in equilibrium. Equilibrium economics then, by its base assumptions, is not primed to look for the exploitation of systems, and as a result systematic studies of how systems might fail or be exploited are not central to how the discipline thinks.[2]

In this paper I want to get away from the equilibrium assumption and take as our basis a different, nonequilibrium assumption: that any policy system at any time presents incentives to the parties engaged in it, and these incentives may in turn induce parties to discover ways in which they might privately benefit that policy designers had not thought of. Given this, we would want to know how exploitive behavior for policy systems might typically arise, and how we can use formal modeling and analysis to allow for such behavior, and to foresee or even warn of it in advance.

I will pose our problem of foreseeing possible exploitation as four questions I will look at in sequence. First, what are the causes of exploitive behavior and how does it typically arise? Second, given a particular economic system or proposed policy, how might we anticipate where it might fail, and what can we learn from disciplines such as structural engineering that try to foresee potential failure modes, and could help us in this? Third, how can we construct models of systems being gamed or exploited, and of agents in these models "discovering" ways to exploit such systems? And fourth, what are the future prospects for constructing artificially intelligent methods that could automatically anticipate how economic and social systems might be exploited? Fully definitive answers to these questions are of course not possible, but I hope the discussion here will at least open the subject for debate.

Before we go on, a word about some of the terms I will use. Exploitation has two meanings: "to use something in order to gain a benefit," and to take "selfish or unfair advantage of a person or situation, usually for personal gain."[3] The first meaning suits us well

[2] There are many critiques of economics in the face of the 2008 financial crisis. E.g. Colander *et al.*, 2008; Koppl and Luther, 2010.

[3] Microsoft Word Dictionary (1991) uses "exploitation" in its first sense, as the use and refinement of existing opportunities, and contrasts this with "exploration," the ongoing search for new opportunities.

(note it is not necessarily pejorative), but the second also covers many of the cases I will talk about. Gaming itself has a more pernicious meaning: it denotes people using a system cynically to their own ends, often in a way that betrays trust placed in them and harms other people.[4] I will also talk of policy systems, meaning economic or social or military or business or governmental systems that play out over time, given a set of policies that define them. The 2010 Obama Affordable Health Care system is a policy system.

Causes of exploitive behavior

Before we talk about modeling exploitive behavior, it will be useful to build up some knowledge about its causes and mechanisms.

Our first observation is that exploitive behavior is not rare. This is not because of some inherent human tendency toward selfish behavior; it is because all policy systems—all social policies—pose incentives that are reacted to by groups of agents acting in their own interest, and often these reactions are unexpected and act counter to the policy's intentions. Examples are legion. The 2003 US invasion of Iraq—a military policy system—was well planned and well executed, but it generated insurgency, a less than fully expected reaction to the presence of American soldiers that went on to obstruct US goals in Iraq. The 1965 Medicare system, launched under Lyndon Johnson with the purpose of providing health care for the elderly, paid fee-for-service, compensating hospitals and physicians for their incurred costs of treatment. Hospitals and physicians in the program responded by purchasing expensive equipment and providing services that were unnecessary. As a result, within five years of its inception, the program's costs nearly tripled (Mahar, 2006). A decade or two later, the United States opened health care to market forces. The freeing of the market was intended to produce competition and to lower costs. Instead it produced a system where each of the key players found specific ways to work the system to their own advantage, to the detriment of the system as a whole. Maher (2006) describes the outcome as "a Hobbesian marketplace" that pitted "the health care industry's players against one another: hospital vs. hospital, doctor vs. hospital, doctor vs. doctor, hospital vs. insurer, insurer vs. hospital, insurer vs. insurer, insurer vs. drugmaker, drugmaker vs. drugmaker."

These examples are large-scale ones, but exploitation happens on every scale. Apartment building managers have been known to visit their competitors' buildings and post negative ratings online to enhance their own competitive standing. Whatever the scale at which exploitation takes place, its frequency of occurrence should give us pause about implementing any social policy without thinking through how it could potentially be used to players' advantage, and it should also caution us about accepting the results of economic models designed to demonstrate a policy system's outcome. In fact, it should caution us about accepting the results of all policy models without questioning their built-in assumptions.

See J. March (1991) on exploitation versus exploration. "Exploitation" in this paper contains elements of both: we are talking of agents exploring for opportunities to exploit.

[4] Wikipedia (October 9, 2010) defines gaming as "[using] the rules and procedures meant to protect a system in order, instead, to manipulate the system for a desired outcome."

But just how should we question the outcome of policy systems? The examples I have given seem scattered and unique, so it doesn't seem easy to build general insights from them. It would be better if we could find generic categories of exploitation, standard hacks, or patterns of behavior or incentives that we see repeated from one circumstance to another. Or to put this another way, it would be useful if we had a "failure mode analysis" tradition in economics for assessing policy systems. Such a tradition exists in other disciplines where life or safety or well-being are at stake: Failure mode analysis in engineering investigates the ways in which structures have failed in the past and might fail or not function as intended; preventive medicine and disease control investigates the causes of diseases, death, and epidemics and looks to their future prevention. These modalities seek not just to study past failures but to construct an organized body of knowledge that might help prevent failure or breakdown in the future.

It would be a large undertaking to construct a policy-system failure mode sub-discipline of economics, worthwhile of course, but beyond the scope of this paper. What we can do is think about how such a discipline would work. One good place to start is to look at how systems have been exploited or gamed in the past and point to general categories or motifs by which this happens. I will talk about four motifs and label these by their causes:

1. Use of asymmetric information In many social systems, different parties have access to different information, and often one party offers a service or puts forward an opportunity based upon its understanding of the available information. Another party then responds with behavior based on its more detailed understanding and uses the system to profit from its privileged information. The financial and the marketing industries are particularly prone to such behavior; in each of these some parties are well informed about the product they are promoting, while others—the potential investors or customers—are not. In 2007 Goldman Sachs created a package of mortgage-linked bonds it sold to its clients. But it allowed a prominent hedge fund manager, John A. Paulson, to select bonds for this that privately he thought would lose value, then to bet against the package. Paulson profited, and so allegedly did Goldman by buying insurance against loss of value of the instrument; but investors lost more than $1 billion (Appleton, 2010). The package (a synthetic collateralized debt obligation tied to the performance of subprime residential mortgage-backed securities) was complicated, and its designers, Goldman and Paulson, were well informed on its prospects. Their clients were not.

The health care industry is also prone to information asymmetries; both physicians and patients are better informed on ailments and their appropriate treatments than are the insurance companies or governmental bodies paying for them (Arrow, 1963). In 2006 the state of Massachusetts mandated individual health care insurance, and the program appeared to work initially, but after some few months insurers discovered they were losing money. The reason was, as Suderman (2010) reports, "[t]housands of consumers are gaming Massachusetts' 2006 health insurance law by buying insurance when they need to cover pricey medical care, such as fertility treatments and knee surgery, and then swiftly dropping coverage." This behavior is not illegal, nor is it quite immoral, but it is certainly exploitive.

2. Tailoring behavior to conform to performance criteria A second type of exploitation—better to call it manipulation here—occurs when agent behavior is judged, monitored, or measured by strict criteria of evaluation and agents optimize their behavior to conform to these narrow criteria, rather than to what was more widely intended. Agents, in other words, game the criteria. Before the 2008 financial crisis, financial ratings agencies such as Moody's or Standard & Poor's for years performed evaluations of the risk inherent in financial instruments proposed by investment and banking houses. A few years before the financial crash, in an act of transparency and implicit trust, they made their ratings models available to the Wall Street investment firms. Says Morgenson (2010): "The Wall Street firms learned how to massage these models, change one or two little inputs and then get a better rating as a result. They learned how to game the rating agency's models so that they could put lesser quality bonds in these portfolios, still get a high rating, and then sell the junk that they might not otherwise have been able to sell."[5]

Gaming performance criteria is not confined to Wall Street. It occurs within all systems where judgment of performance is important: conformance to the law; educational testing[6]; adherence to standards of human rights; adherence to environmental standards; adherence to criteria for receiving funding; the production of output within factories; financial accounting; tax reporting; the performance of bureaucrats; the performance of governments. In all these cases, the parties under surveillance adjust their behavior to appear virtuous under the stated performance measures, while their actual behavior may be anywhere from satisfactory to reprehensible. In fact, in the case of government performance, two particular expressions of this form of exploitation already exist. One is Campbell's law (1976): "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intend to monitor." The other is Goodhart's law (1975)[7]: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." Both of these apply to governmental behavior. I prefer a broader truism: Any performance criterion will be optimized against, and will thereby lose its value.

3. Taking partial control of a system A third type of exploitation occurs when a small group of agents manages to take control of some significant portion of the resources of a system and use this for its own purposes. This is the economic equivalent of the way viruses operate. The group in effect takes over part of the machinery of the system and uses that to its own advantage.

The financial sector has seen a great deal of this type of exploitation. Within the insurance giant AIG, some years before the 2008 crash, a small group of people (the Financial Products Unit) managed to take effective control of much of the company's assets and risk bearing, and began to invest heavily in credit default swaps. The group profited greatly through their own personal compensation—they were paid a third of the profits they generated—but the investments collapsed, and that in turn sank AIG (Zuill, 2009). A similar set of events unfolded in Iceland, where a small group of entrepreneurs took out loans, used these to buy

[5] See (White (2009).

[6] See Nichols and Berliner (2007).

[7] Chrystal and Mizen (2001).

control of the assets of the country's banks, and invested these in international properties and derivatives (Boyes, 2009; Jonsson, 2009). The international investments collapsed, and so did Iceland's banks along with their customers' deposits.

4. Using system elements in a way not intended by policy designers Still another type of exploitation happens when agents use the behavior of the system itself to manipulate the system. An example would be using a website's rating possibilities to manipulate others' ratings. Often too, players find a rule they can use as a loophole to justify behavior the designers of the system did not intend. Usually this forces a flow of money or energy through the rule, to the detriment of the system at large. Following the Arab Oil Embargo in the early 1970s, the US Congress set up fuel economy standards for motor vehicles. Understandably, the requirements for commercial light trucks were more lenient than those for passenger vehicles. But in due course and with a little congressional manipulation, Detroit found it could declare its sports utility vehicles to be light trucks. These then passed through the light-truck loophole, the highways in due course filled with SUVs, and between 1988 and 2005 average fuel economy actually fell in the United States (Pew, 2010). This was not what the energy policy's designers intended.

The four motifs I have described are by no means exhaustive; there are no doubt other ways in which systems might be gamed. But these give us a feel for the types of exploitation we might expect to see, and they show us that exploitive behavior is not rare in systems. It is rife.

Anticipating failure modes

For some policy systems, it is obvious that their possible exploitation falls into one of the four motifs just given. For others no particular mode of exploitive behavior might be obvious. In general we have a given policy system and a mental model or analytical studies of how it is expected to work, and we would like to anticipate where the system might in real life be exploited. So how do we proceed in general? How would we go about failure mode analysis in a particular economic situation? There is no prescribed answer to these questions, but we can usefully borrow some directives from engineering failure analysis.

An obvious first step is to have at hand knowledge of how similar systems have failed in the past. We have at least the beginnings of such knowledge with the motifs I described earlier. Aircraft designers know from forensic studies the causes by which failures (they call these "anomalies") typically occur: fatigue failure, explosive decompression, fire and explosions, burst engines (Bibel, 2008). By analogy, as I said, we need a failure mode analysis of how policy systems have been exploited in the past.

Second, we can observe that in general the breakdown of a structure starts at a more micro level than that of its overall design. Breakdown in engineering designs happens not because the overall structure gives way, but because stresses cause hairline cracks in some part of an assembly, or some component assembly fails, and these malfunctions propagate to higher levels, possibly to cause eventual whole-system degradation. This suggests in our case that for any system we are studying, exploitive behavior will typically take place at a smaller scale than the overall system. Exploitive behavior after all is created—is "invented"—by individual people, individual human agents, or small groups of these, and we will have to have detailed knowledge of the options and possibilities agents possess if we want to understand how this may happen.

Third, and again by analogy, we can look for places of high "stress" in the proposed system and concentrate our attentions there. In social systems these places tend to be the points that present strong incentives for agents to do something different from their prescribed behavior. Typically, in an analytical model, points of behavioral action are represented as rates (the rate, say, at which individuals buy health insurance), or as simple rules (if income exceeds $X, and age exceeds Y, buy health insurance). The modeler needs to query whether simple rates or rules are warranted, given the pattern of incentives agents faces. Very often they are not.

All this would suggest that if we have a design for social system and an analytical model of it, we can "stress test" it by first identifying where actual incentives would yield strong inducements for agents to engage in behavior different from the assumed behavior. These might, to give some examples, be places where agents have power to affect other players' well-being (they can issue building permits, say, to wealthy property developers), yet we assume they make impartial decisions; or places where agents can profit by compromising on performance or safety of some activity (say, they decide on aircraft maintenance), yet we assume they conform to given standards; or places where agents have inside information (say, they have knowledge of a company's future plans), yet we assume they do not trade on this information.

Next we construct the agents' possibilities from our sense of the detailed incentives and information the agents have at this location. That is, we construct detailed strategic options for the agents. The key word here is "detailed": the options or opportunities possible here are driven by the imagination and experience of the analyst looking at the system, they are drawn from the real world, and they require careful, detailed description. This is why we will need to have knowledge of the fine-grained information and opportunities the agents will draw from to create their actions.

Once we have identified where and how exploitation might take place, we can break open the overall economic model of the policy system at this location, and insert a module that "injects" the behavior we have in mind. We now have a particular type of exploitation in mind, and a working model of it that we can use to study what difference the strategic agents make in the behavior of the overall system. Sometimes they will make little difference; the strategic behavior may not affect much outside its sphere. Sometimes they will have a major effect; they may even in certain cases cause the collapse of the structure they were inserted into. What is important here is that we are looking for weak points in a policy system and the consequences that might follow from particular behaviors that system might be prone to. It is important that this testing not be rushed. In engineering it often takes months or years to painstakingly test, debug, and rework a novel design of importance, especially where public safety is at stake. There is no reason we should place less emphasis on the safety of economic and social policy outcomes.

This method I have just outlined presumes one system designer, or a team, working to discover flaws in a given set of policies or given simulated economic system. Things can be speeded up if multiple designers work in parallel and are invited to probe a model to find its weak points. Where we have a working model of a proposed policy system—think of a new health care policy, or an altered set of financial regulations—we can solicit "strategy" modules that exploit it. Here the overall simulation model or overall policy situation would be given, and we would be inviting outside participants to submit strategies to exploit it. This was first carried out in the famous prisoner's dilemma tournament several decades ago, where Robert Axelrod (1984) solicited strategies that would compete in a repeated prisoner's dilemma game. To do this in the more general systems context, participants would need to study the system thoroughly, identify its myriad incentives, home in on the places were it leaves open opportunities for exploitation, and model these.

Something similar to this is carried out routinely in the beta testing of encryption systems. When, say, the US Navy develops a novel encryption scheme, it invites a group of selected people to see if they can crack the scheme. If they cannot, the scheme can proceed. It is important that testers come from the outside. Says Schneier (1999): "Consider the Internet IP security protocol. It was designed in the open by committee and was the subject of considerable public scrutiny from the start. … Cryptographers at the Naval Research Laboratory recently discovered a minor implementation flaw. The work continues, in public, by anyone and everyone who is interested. On the other hand, Microsoft developed its own Point-to-Point Tunneling Protocol (PPTP) to do much the same thing. They invented their own authentication protocol, their own hash functions, and their own key-generation algorithm. Every one of these items was badly flawed. … But since they did all this work internally, no one knew that their PPTP was weak."

Modeling exploitation within computer models

In the previous section I talked in general about probing policy systems for possible failure. Now I want to narrow this and talk more about probing computer-based models of policy systems—usually simulation models—for possible failure. One difficulty we immediately face is that most computer-based models are closed to novel behavior: they use equations or Markov states or other architectures that assume fixed categories of behavior laid down in advance or embedded within them, so they can't easily be modified to conjure up the unforeseen—the 51-foot ladders that might appear.

But we can proceed. Certainly, as I said before, we can "inject" foreseen exploitive behavior into the computer model; that's a matter of breaking open the model and adding more detail. More generally, though, we would like to be able to have our simulation model allow for the spontaneous arising or "discovery" of unforeseen novel behaviors, and this seems more challenging. Notice we are really asking how new behaviors might emerge from agents' discovering or learning within a system, and emergence is something that modeling, especially agent-based modeling, has experience with. So we might expect that we can indeed modify a simulation model to allow agents to "discover" manipulative behavior.

Let me illustrate with a real-world example. Consider the health insurance case I mentioned from Massachusetts. We don't have a simulation model of this policy system at hand, so for our purposes we will construct one. And because we are interested not in social details but in issues of how we can simulate exploitation, we can keep this simple and stylized.

We will proceed in steps by constructing versions that progressively capture the behavior that interests us.

First we construct a basic model of health insurance. (I used NetLogo for this simulation, a convenient platform for agent-based modeling.) The model has N (typically from 100 to 1,000) people who individually and randomly incur health care costs, perhaps from diseases, hospital care, surgical procedures, or accidents, and initially they cover these costs themselves. In this model, the distribution of health costs is uniform, stationary, and identical for all (we can assume people are all of the same age). People receive a fixed income, common to all, and their consumption c equals this less their health costs. I assume a concave utility function over consumption, $U(c) = c1/2$: people are risk averse. There is one insurance company. At first it offers no policies, but instead for a fixed period collects actuarial data: It has access to the population's health costs and uses these to figure average health costs per person per period. Once it has a sufficiently accurate estimate it issues a voluntary health insurance policy. The policy's cost is set to be "fair" (equal to its estimate of the expected cost per person per period) plus a markup of m% to cover administrative costs. When we run the model we find that when insurance markup values are sufficiently low ($m < 23.3\%$) people find that their utility is higher with the policy and they buy it. Otherwise they do not. We now have a simple working agent-based model of insurance.

As a second step, let us build in the Massachusetts edict and its consequences. Central to what happened was a class of people who believed or found out they could do without coverage; instead they could pay the fine for non-participation in the scheme. There are several ways we could modify our model to allow for such a class. We could assume, for example, people who have small risk of incurring health costs. But the simplest way is to assume that a proportion of the population (let us say 50%) is not risk-averse. It has a linear utility function, $U(c) = c$, and thus finds it profitable to pay the government fine, assuming (as is true in Massachusetts) this is less than the insurance markup. When we run this model we find not surprisingly that one half of the population insures, the other half does not.

As a third step we build in the exploitive behavior. We now allow that all people can see costs in advance for some types of health care (shoulder operations, say, or physical therapy). So we build into the model—"inject" the behavior—that these can be foreseen at the start of the period, giving people the option of taking out coverage for that period and possibly canceling it the next. The 50% already insured will not be affected; they are paying insurance regardless. But the uninsured will be affected, and we find when we run this model that they opt in and out of coverage according to whether this suits their pockets. In the sense that they are taking out insurance on an outcome they know in advance, but the insurance company does not, they are "gaming" the system. Figure 1 shows the consequences for the insurance company's profits when it switches in. They plummet.

As a last stage in our modeling, realistically, we can assume that the system responds. The state may raise its non-participation fine. Once it does this sufficiently we find that everyone participates and normality resumes. Or the insurance company can react by increasing the mandatory policy-holding period. Once it does this to a sufficient point, we find again that normality resumes.

I have constructed this model in stages because it is convenient to break out the base model, demarcate the agents that will strategize, allow them to do so, and build in natural responses of the other agents. When finished, the model runs through all these dynamics in sequence, of course.

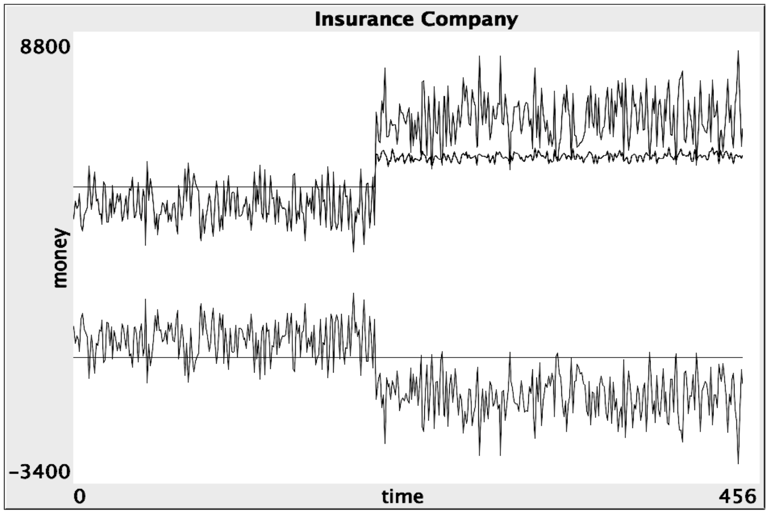

Fig 1. Agents are allowed to see some upcoming health expenses starting around time 300. The upper plots show the effect on the insurance company's income from policy payments (smoother line) which rises because it acquires additional policy holders, and from its expenses (upper jagged line) which rise due to preplanned claims. The lower plot shows the company's profits (lower jagged line), which now fall below the flat zero line.

So far this demonstrates that we can take a given simulation model of a policy system (we constructed this one) and modify it by injecting foreseen "exploitive" behavior and response into it. But, as I mentioned earlier, in real life, exploitation emerges: it arises—seemingly appears—in the course of a policy system taking its course. In fact, if we look at what happens in real life more closely, we see that players notice that certain options are available to them, and they learn from this—or sometimes discover quite abruptly—that certain actions can be profitably taken. So let us see how we can build "noticing" and "discovery" in to our example. To keep things short I will only briefly indicate how to do this.[8]

First, "noticing" is fairly straightforward. We can allow our agents to "notice" certain things—what happened say in the recent past, what options are possible—simply by making these part of the information set they are aware of as they become available (cf. Lindgren, 1992).

[8] See Arthur *et al.* (1997) for a study that implements the procedure I describe.

We still need to include "discovery." To do this we allow agents to generate and try out a variety of potential actions or strategies based on their information. There are many ways to do this (Holland, 1975; Holland et al, 1986). Agents can randomly generate contingent actions or rules of the type: If the system fulfills a certain condition K then execute strategy G. Or they can construct novel actions randomly from time to time by forming combinations of ones that have worked before: If conditions K and P are true, then execute strategy F. Or they can generate families of possible actions: Buy in if this period's pre-known health costs exceed k dollars (where k can be pegged at different levels). We further allow agents to keep these potential strategies in mind (there may be many) and monitor each one's putative performance, thus learning over time which ones are effective in what circumstances They can then use or execute the strategy they deem most effective at any time; and drop strategies that prove ineffective.

This sort of design will bring in the effect we seek. (For detailed illustrations of it in action, see Arthur, 1994, and Arthur et al., 1997.) If some randomly generated strategy is monitored and proves particularly effective, certain agents will quickly "discover" it. To an outsider it will look as if the strategy has suddenly been switched on—it will suddenly emerge and have an effect. In reality, the agents are merely inductively probing the system to find out what works, thereby at random times "discovering" effective strategies that take advantage of the system. Exploitation thus "appears."

I have described a rather simple model in our example and sketched a way to build the emergence of possible exploitations into it. Obviously we could elaborate this in several directions.[9] But I want to emphasize my main point. Nothing special needs to be added by way of "scheming" or "exploitive thinking" to agent-based simulation models when we want to model system manipulation. Agents are faced with particular information about the system and the options available to them when it becomes available, and from these they generate putative actions. From time to time they discover particularly effective ones and "exploitation"—if we want to call it that—emerges. Modeling this calls only for standard procedures already available in agent-based modeling.

Automatic pre-discovery of exploitive behavior

But this is still not the last word. In the previous section, if we wanted computation to "discover" exploitive behaviors as in the previous section, we needed to specify a given class of behaviors within which they could explore. Ideally, in the future, we would want computation to automatically "discover" a wide range of gaming possibilities that we hadn't thought of, and to test these out, and thereby anticipate possible manipulation.

[9] For example, realistically we could allow information on what works to be shared among agents and spread through their population, and we could assume that if a strategy works, people would focus attention on it and construct variants on it—they would "explore" around it.

What are the prospects for this? What would it take for a simulation model of US-Mexico border crossings to foresee—to be able to "imagine"—the use of 51-foot ladders? Of course, it would be trivially easy to prompt the computer to see such solutions. We could easily feed the simulation the option of ladders of varying length, 40-foot, 64-foot, 25-foot, and allow it to learn that 51-foot ladders would do the job just right. But that would be cheating.

What we really want, in this case, is to have the computer proceed completely without human prompting, to "ponder" the problem of border crossing in the face of a wall, and "discover" the category of ladders or invent some other plausible way to defeat obstacles, without these being built in. To do this the computer would need to have knowledge of the world, and this would have to be a deep knowledge. It would have to be a general intelligence that would know the world's possibilities, know what is available, what is "out there" in general outside itself. It would need in other words something like our human general intelligence. There is more than a whiff of artificial intelligence here. We are really asking for an "invention machine": a machine that is aware of its world and can together put available components conceptually to solve general problems. Seen this way, the problem joins the category of computational problems that humans find doable but machines find difficult, the so-called "AI-complete" problems, such as reading and understanding text, interpreting speech, translating languages, recognizing visual objects, playing chess, judging legal cases. To this we can add: imagining solutions.

It is good to recognize that the problem here is not so much a conceptual one as a practical one. We can teach computers to recognize contexts, and to build up a huge store of general worldly knowledge. In fact, as is well known, in 2010 IBM taught a computer to successfully answer questions in the quiz show Jeopardy, precisely through building a huge store of general worldly knowledge. So it is not a far cry from this to foresee computers that have semantic knowledge of a gigantic library of past situations and how they have been exploited in the past, so that it can "recognize" analogies and use them for the purpose at hand. In the case of the 2003 US invasion of Iraq, such computation or simulation would have run through previous invasions in history, and would have encountered previous insurgencies that followed from them, and would have warned of such a possibility in the future of Iraq, and built the possibility into the simulation. It would have anticipated the "emergent" behavior. Future simulations may well be able to dip into history, find analogies there—find the overall category of ladders as responses to walls—and display them. But even if it is conceptually feasible, I believe full use of this type of worldly machine intelligence still lies decades in the future.

Conclusion

Over the last hundred years or more, economics has improved greatly in its ability to stabilize macro-economic outcomes, design international trade policies, regulate currency systems, implement central banking, and execute antitrust policy. What it hasn't been able to do is prevent financial and economic crises, most of which are caused by exploitive behavior. This seems an anomaly given our times. Airline safety, building safety, seismic safety, food and drug safety, disease safety, surgical safety—all these have improved steadily decade by decade in the last fifty years. "Economic safety" by contrast has not improved in the last five decades; if anything it has got worse.

Many economists—myself included—would say that unwarranted faith in the ability of free markets to regulate themselves bears much of the blame (e.g. Cassidy, 2009; Tabb, 2012). But so too does the absence of a systematic methodology in economics of looking for possible failure modes in advance of policy implementation. Failure-mode studies are not at the center of our discipline for the simple reason that economics' adherence to equilibrium analysis assumes that the system quickly settles to a place where no agent has an incentive to diverge from its present behavior, and so exploitive behavior cannot happen. We therefore tend to design policies and construct simulations of their outcomes without sufficiently probing the robustness of their behavioral assumptions, and without identifying where they might fail because of systemic exploitation.

I suggest that it is time to revise our thinking on this. It is no longer enough to design a policy system and analyze it and even carefully simulate its outcome. We need to see social and economic systems not as a set of behaviors that have no motivation to change, but as a web of incentives that always induce further behavior, always invite further strategies, always cause the system to change. We need to emulate what is routine in structural engineering, or in epidemiology, or in encryption, and anticipate where the systems we study might be exploited. We need to stress test our policy designs, to find their weak points and see if we can "break" them. Such failure-mode analysis in engineering, carried out over decades, has given us aircraft that fly millions of passenger-miles without mishap and high-rise buildings that do not collapse in earthquakes. Such exploitation-mode analysis, applied to the world of policy, would give us economic and social outcomes that perform as hoped for, something that would avert much misery in the world.

References

Appleton, Michael, "SEC Sues Goldman over Housing Market Deal." *New York Times*, Apr 16, 2010.

Arrow, Kenneth, "Uncertainty and the Welfare Economics of Medical Care," *American Econ. Rev* 53: 91-96, 1963.

Arthur, W. Brian. "Bounded Rationality and Inductive Behavior (the El Farol problem)," *American Economic Review Papers and Proceedings*, 84, 406-411, 1994.

Arthur, W. Brian, J.H. Holland, B. LeBaron, R. Palmer, and P. Tayler, "Asset Pricing under Endogenous Expectations in an Artificial Stock Market, in *The Economy as an Evolving Complex System II*, Arthur, W.B., Durlauf, S., Lane, D., eds. Addison-Wesley, Redwood City, CA, 1997.

Axelrod, Robert. *The Evolution of Cooperation*. Basic Books, New York. 1984.

Bibel, George. *Beyond the Black Box: The Forensics of Airplane Crashes*, Johns Hopkins Univ. Press, Baltimore, MD, 2008.

Boyes, Roger. *Meltdown Iceland*: *How the Global Financial Crisis Bankrupted an Entire Country*. Bloomsbury Publishing, London. 2009.

Campbell, Donald, "Assessing the Impact of Planned Social Change," Public Affairs Center, Dartmouth, NH, Dec, 1976.

Cassidy, J., *How Markets Fail*: *the Logic of Economic Calamities*, Farrar, Straus and Giroux, NY, 2009.

Chrystal, K. Alec, and Paul Mizen, "Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy," (http://cyberlibris.typepad.com/blog/files/Goodharts_Law.pdf), 2001.

Colander, David, A. Haas, K. Juselius, T. Lux, H. Föllmer, M. Goldberg, A. Kirman, B. Sloth, "The Financial Crisis and the Systemic Failure of Academic Economics," mimeo, 98th Dahlem Workshop, 2008.

Holland, John. *Adaptation in Natural and Artificial Systems*. MIT press. 1992. (Originally published 1975.)

Holland, John H., K. J. Holyoak, R. E. Nisbett and P. R. Thagard, *Induction*. Cambridge, MA: MIT Press. 1986.

Jonsson, Asgeir. *Why Iceland? How One of the World's Smallest Countries became the Meltdown's biggest Casualty,* Mc-Graw-Hill, New York. 2009.

Koppl, Roger, and W.J. Luther, "BRACE for a new Interventionist Economics," mimeo, Fairleigh Dickinson Univ., 2010.

Lindgren, Kristian. "Evolutionary Phenomena in Simple Dynamics," in C. Langton, C. Taylor, D. Farmer, S. Rasmussen, (eds.), *Artificial Life II*. Addison-Wesley, Reading, MA, 1992.

Mahar, Maggie. *Money-driven Medicine*. HarperCollins, New York. 2006.

March, James, "Exploration and Exploitation in Organizational Learning," *Organization Science*, 2, 1, 71-87, 1991.

Morgenson, Gretchen, (New York Times reporter), "Examining Goldman Sachs," NPR interview in *Fresh Air*, May 4, 2010.

Nichols, S. L. and D. Berlner, *Collateral Damage: How High-stakes Testing Corrupts America's Schools.* Harvard Education Press, Cambridge, Ma. 2007.

Pew Charitable Trusts, "History of Fuel Economy: One Decade of Innovation, Two Decades of Inaction," www.pewtrusts.org., 2010.

Schneier, Bruce, "Cryptography: The Importance of Not Being Different," *IEEE Computer*, 32, 3, 108-109, 1999.

Suderman, Peter, "Quit playing Games with my Health Care System," *Reason*, April 5, 2010.

Tabb, William, *The Restructuring of Capitalism in our Time*, Columbia University Press, New York, 2012.

White, Lawrence J. "The Credit Rating Agencies and the Subprime Debacle." Critical Review, 21 (2-3): 389-399, 2009

Zuill, Lilla, "AIG's Meltdown Has Roots in Greenberg Era." Insurance Journal, 87, March 3, 2009.