Ministry of Foreign Affairs of the Netherlands

Radboud University

Special Working Paper Series on ‘Unintended Effects of International Cooperation’

Working Paper No. 2017/10

From Firefighting to Systematic Action:

Toward A Research Agenda for Better Evaluation of Unintended Consequences

Jonathan A. Morell *

Abstract

This paper is an attempt to improve the ability of evaluation to detect and measure unintended consequences of program action. One audience is the evaluation community, which needs new approaches and methodologies. A second audience is the donor community, which must alter the demands it makes on evaluators, and which should also sponsor research on improving evaluation methods. Evaluators would benefit from engaging in continually iterating between their models and their empirical data. There are no restrictions on the form or detail in the models, as long as the iteration takes place systematically over the entire evaluation lifecycle. Diverse input is needed when data are used to revise models. Efficient and effective methods are proposed to elicit that feedback while avoiding the bias and conflict that comes from face to face group interaction. A research agenda is offered to explore the value of the proposed changes in evaluation practice. Specific questions to be researched are identified, and methodologies to address these questions are sketched. Both evaluators and donors should appreciate why program outcomes may be unpredictable, how change happens, and why change is hard to detect. Relevant issues are the continuum between predictable and unpredictable change, incremental and discontinuous change trajectories, network effects, feedback loops, the relationship between system behavior and the social/economic/political drivers of program design, and the complex settings in which programs are embedded. Policy recommendations are presented that are designed to improve the content of models that drive evaluations, and to fund necessary research into developing evaluation methods that are better able to address unintended outcomes of program action.

Key words: Unintended consequences, evaluation, program models, types of change, complex systems

* Ph.D. 4.669 Evaluation and Planning Editor, Evaluation and Program Planning

This paper has been presented at the ‘Unintended Effects of International Cooperation’ conference on January 16&17, 2017 at the Netherlands Ministry of Foreign Affairs in The Hague. The opinions expressed in this article are the author's own and do not reflect the view of neither the Netherlands Ministry of Foreign Affairs, nor the Radboud University, or any other affiliated institutions.

As a program evaluator my interest is in the consequences of program action. *Can an apparent consequence of program action really be attributed to that program?*

It is hard enough to answer this question when we have time to define what outcomes we want to measure, and to devise methodology that is trained on those outcomes. It is harder when unintended outcomes occur because then we need to change a methodology that is already in place. (Following the lead of the literature review that led to this conference, I use the term "unintended" to refer to consequences that may be desirable or undesirable, anticipated or unanticipated (Koch & Schulpen, 2017).

We evaluators always manage to muddle through and come up with something when surprises pop up. But a motivating theme in much of my work is that we can do better. We can implement processes that will either anticipate surprise, or respond in a powerful way when surprise appears. We do not always have to deal with crisis. We can anticipate and plan. We can include planning for unintended consequences along with all the other considerations that go into designing an evaluation. We can move from firefighting to systematic action (Jonathan A. Morell, 2005, 2010).

Models and Methodology

Because of my evaluation interests, I have an intense concern with models and with methodology.

- Models are needed for the same reason they are needed in all research; to highlight relationships we care about by deliberately distorting or eliminating others.
- Methodology is needed to uncover empirical data about the models and about the consequences of their operation.

The Relationship Between Model and Methodology

The relationship is mutual. Models drive methodology, while the data yielded from methodology informs the models. There is nothing innovative about this formulation, but I hope to show how the model/data relationship can form the basis of a research agenda to improve the evaluation of unintended consequences of program action. I am going to propose that we systematize a process that many of us do intuitively, albeit haphazardly or irregularly.

My colleagues and I have been researching the value of integrating agent-based simulation with traditional evaluation over the life cycle of an evaluation. Our hypothesis is that by iterating empirical data collection and agent based modeling, we would provide better guidance to policy makers and program planners. It is true that agent-based modeling is interesting and powerful, but equally important in our work is the simple notion of being systematic. Continually feed empirical evaluation data to the model builders. Use the models to provide insight for further data collection (J. A. Morell, et. al., 2016a, 2016b; J. A. Morell, et.al, 2016; J.A. Morell, Hilscher, Magura, & Ford, 2010; Parunak & Morell, 2014).

Our success led me to a eureka moment – if it can work for computer modeling, why shouldn't it work for *any* modeling? It should not matter whether the models qualitative or quantitative, detailed or sparse, with or without embedded levels of granularity, with or without feedback loops, short or long term, with without precise relationships among its elements, with or without probability relationships, or living inside or outside of a computer. What matters is the systematic interplay of model and methodology. The form of the model is irrelevant. It may be as simple as a few lists of words in columns on a piece of paper, or be precise and elaborate. If the model can guide good program evaluation, that's all that counts.

Models in Support of Methodology

The theme of “guiding good evaluation methodology” is what I care about. Besides evaluation, there are other good reasons to formulate models, e.g. to advance advocacy for a program or to explain why a program will work as it does, or why it will have the impact that is expected. Those are legitimate reasons for developing a model, but for the purposes of evaluation, such models tend to have too little, or too much detail to be useful as a practical guide to evaluation design and data interpretation. To my evaluator’s sensibilities, the main reason for having a model is for guidance on methodology. It has to work as a technology. It does not have to be true, but it does have to be useful for the hands-on, real world tasks of collecting data to understand program impact. (For a fuller argument advocating for evaluation as technology rather than science, see: (Jonathan A Morell, 1979)

What are the characteristics of useful models?

1. The model should convey a qualitative sense of change, e.g.
   - Is change continuous or discontinuous, sudden or gradual?
   - Will outcomes appear immediately after a program is implemented, or only after some time has elapsed?

2. The model should not mix levels of detail without good reason
   - As an example, imagine a program to improve cooperation between one group of programs aimed at improving agricultural production, and another designed to support export growth. The theory is that by coordinating formal meetings among the groups, the richness of cooperation among them would increase.

     Should the evaluation include a measure of attitude change that came about as a result of the *particular way* the meetings were conducted? Including that measure would imply a belief that a particular method of running the meetings was necessary for facilitating cooperation. Do we believe that? Even if we did, does it matter? If the answers to these questions is “yes”, then we should include the measure and the methodology it would require. If not, why work with a model that implicitly assumed the importance of that lower level theory, and in the bargain added cost and complexity to the evaluation?

3. The model should be honest with respect to what we do and do not know about the program e.g.

- Do we really believe that the single causal path that we have traced among outcomes is the only one that may occur?
- Do we really believe that all those 1:1, 1:many, many:1, many:many relationships that we have specified are actually how things work?
- Are we sure that those intermediate outcomes in the model are really needed for the program to produce the results we want it to?

Models may be indispensable, but they can also be fickle. They may be accurate enough for an evaluation’s needs over the entire lifecycle on an evaluation, or they can quickly, and often in invisible ways, be overtaken by events. Usually, models do not predict very well (Orrell, 2007). Why? Because:

- Environments change.
- Programs are not stable.
- Stakeholder needs change
- Models are simplifications of reality
- Small events or inaccuracies can result in big changes

Further, even when the predictions are acceptable, they will exclude domains of knowledge and understanding that may be responsible for unintended outcomes. This is because whenever we construct a model, we commit an act of willful ignorance (Weisberg, 2014). To bring forth patterns and relationships that we wish to understand, we deliberately omit others.

Our suppositions about what should be included and excluded in a model may well change over time and with the availability of data. Hence the need for systematic interplay of data and model over the course of an entire evaluation. Either the model may be proved wrong, or the useful domains of ignorance may have shifted. Or both. Continual updating based on the most current data is the only way to assure that models will serve as guides to the future of our evaluation design, what data we collect, and how we interpret that data. Figure 1 is a schematic view of what I have in mind.

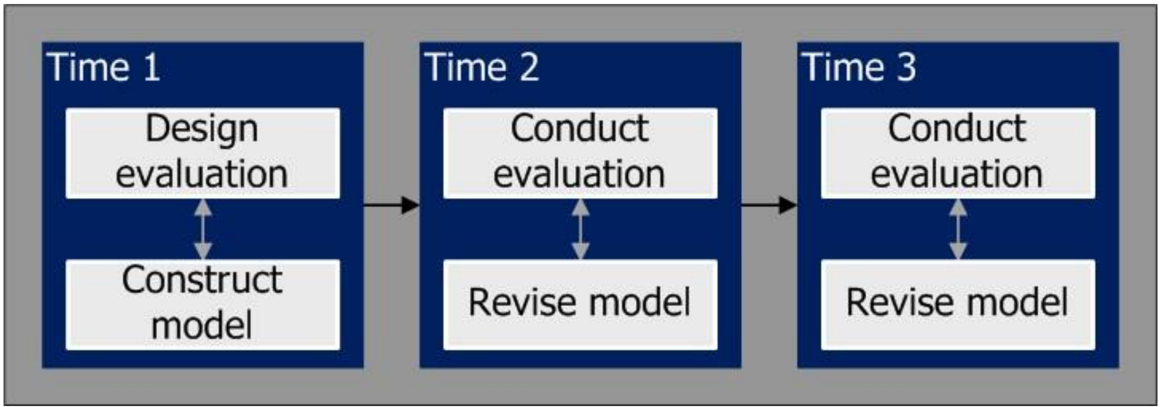

Figure 1: Interplay of Model and Data

What does the relationship between model and methodology look like? At first blush, it seems as if model building is a prelude to methodology, as if model and methodology are engaged in a dialogue.

Dialogue at the start of the evaluation

Model:	Methodology
The program will affect X.	
	I can’t do a very good job of measuring X. But I can measure something related, X**’**
I had not thought to include that construct in my calculations, but I will now.	

This dialogue takes place at the beginning of an evaluation. My contention is that as the evaluation proceeds, another type of dialogue needs to take place.

Dialogue as the evaluation proceeds

Model:	Methodology
	You told me to look for X. I did not find it, but I did find Y.
Ah. I never thought that would be an outcome. I’ll incorporate it into my next run.	
.	After you incorporate Y and look at your output, let me know if that brings up any other things you want me to collect data on.

I do not mean to imply that this latter dialogue does not take place in current evaluation practice. It does. We all do it. But we do not do it systematically. Most important, we do not think about what it means to do it at useful intervals. That consideration is not standard operating procedure in evaluation practice. What does “useful interval” mean? It means intervals that give the evaluation sufficient lead time to develop a design that will address whatever new measurement is required. Also, I do not want to imply that the kernel of my idea is novel in social science. On the contrary, there is a long history of thought on the interplay of different methodologies (Mertens et al., 2016).

A model should be built as soon as one has a reasonable idea of what to model, and no sooner. This is usually when programs are first designed, but it does not always have to be so. My favorite example is the case of introducing mobile technology for telecommunications. Beyond the facile notion that “people will have richer communication”, it is really impossible to foresee how communications may affect banking, family relationships, commercial relationships, civil society, and so on. On the other hand, once those effects become visible, one needs a good model to do a good evaluation of their consequences.

“Lead Time” in Evaluation Redesign

By “lead time” I mean the time from when an unintended consequence is first suspected, to the time when evaluation is capable of assessing that unintended consequence. Putting in effort to assure adequate lead time would not be worth the investment if an evaluation design: 1) relied primarily on post-test interviewing of service recipients who were already involved in the evaluation, and 2) had no need for any kind of comparison group. So what if a need popped up to ask a few different questions? Just ask. The difficulty is that this is a very restrictive example. It limits the use of a large number of methodologies. To illustrate, I’ll lay out a somewhat elaborate example.

Example: Lead Time – Original Program

We have a program that is designed to enrich civil society by building relationships among a variety of different civil society groups in country X. The unit of analysis is “country region”. The program is designed to last for two years. The program itself is simple. Set up formal meetings among the civil society organizations, and give them a chance to interact. Also provide technical assistance regarding expertise and knowledge they may need Figure 2.

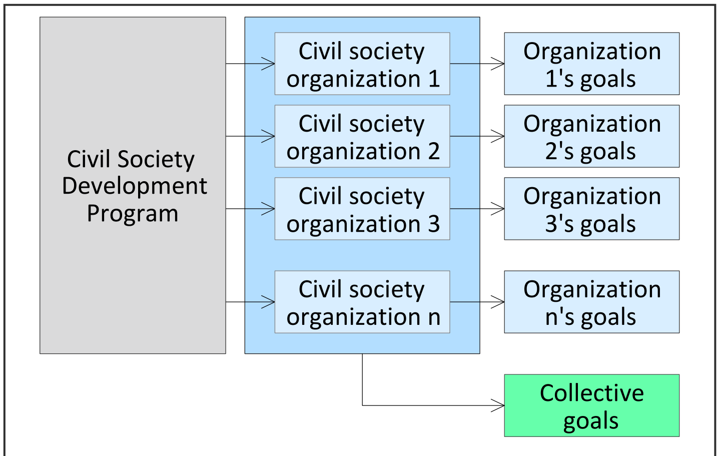
Civil Society Development Program
Civil society organization 1
Civil society organization 2
Civil society organization 3
Civil society organization n
Organization 1's goals
Organization 2's goals
Organization 3's goals
Organization n's goals
Collective goals

Figure2: Civil Society- Original Design

The hypothesis is that as connections among civil society organizations grow, two outcomes will be seen. First, each individual organization will become more effective. Second, a network of organizations will coalesce that will be able to undertake joint action that none alone could accomplish. To be useful as a guide to evaluation, Figure 2 would have to be fleshed out with a few intermediate outcomes, relationships among them, and perhaps, a few critical feedback loops. But for all the elaboration, the overall logic would remain the same.

Outcome measures are: 1) Indicators of group action, e.g. How many meetings did they hold? 2) What interactions do they have with various groups, e.g. the business community, government, the educational establishment, etc.? 3) What changes take place in the programs and services they provide? 4) How do recipients of the civil society action perceive the effectiveness of the outreach?

Because many of these actions are informal and not documented, regular interviews are conducted. Otherwise, the ephemeral knowledge would be lost. Another outcome measure uses a validated survey scale to measure people’s belief in the effectiveness of their organization. The survey compares baseline with beliefs after year, and then again at the end of the program. There is no control group because we are sure that documenting activity within the existing organizations will provide convincing evidence as to whether the program was or was not effective.

Example Lead Time – Evolved Program

Imagine that well into the program, we suspect that another outcome is occurring. Namely, that entirely new civil society groups are forming. We surmise that because of all the activity that our program is producing, people in the community become interested in forming new organizations. This outcome is different from existing groups becoming more effective, or new levels of cooperation among those existing groups.

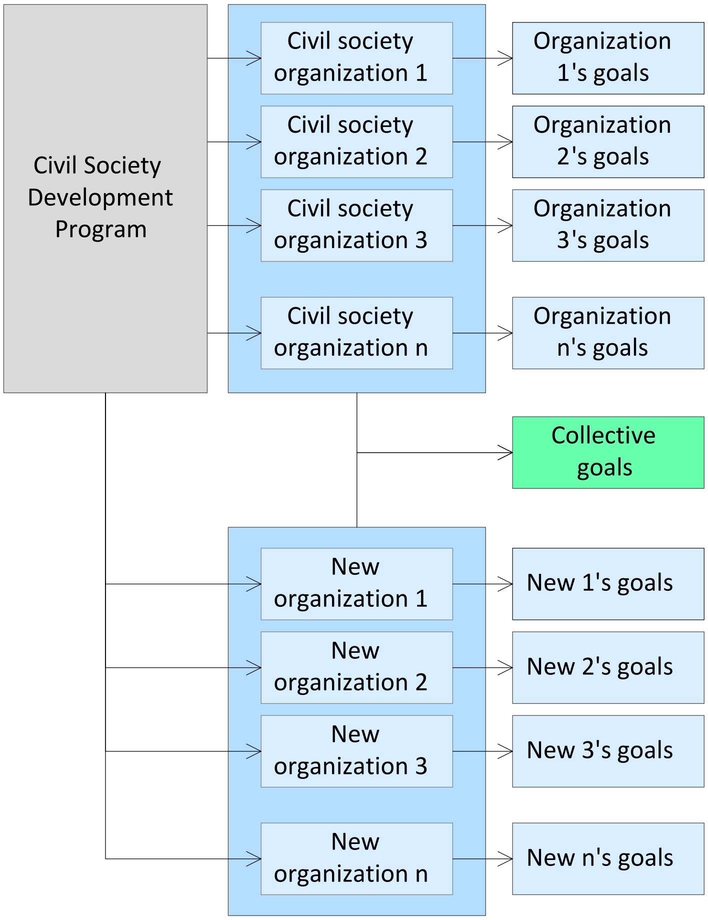

Figure 3: Civil Society Evolved Design

Is the formation of new groups a function of our program? That is certainly possible, but then again, the formation of civil society groups might be a function of a great many factors that are completely external to our program. We think our program model has evolved in a profound way, as illustrated in Figure 3. But we are not sure, and we need an evaluation design to find out.

Addressing the evaluation question depicted in Figure 3 would be exceedingly difficult with a design that was wrapped around the model in Figure 2. How so? Because of the need for evaluation elements that we had not planned for. For instance, we might want to look at regions where our program was not operating in order to determine the natural evolution of civil society groups. We might want to conduct interviews or administer the "perceived effectiveness" scale to groups of people who were not already familiar and accepting of us. We might want to study the purposes of thnewly formed organizations to see if there is overlap with the existing civil society infrastructure. We might want to determine the extent of joint memberships among the old and new organizations. We might want to assess the consequences of new organizations trying to impact the same governmental processes that are the target of the existing organizations.

In the above example "lead time" becomes critical because would need time to:

- establish relationships with the new groups,
- find some new domain experts and consultants,
- renegotiate the terms of reference of the evaluation,
- gear up data collection in regions where we had not worked before, and
- figure out a way to find out if our networking efforts had anything to do with new group formation

All this would require a lot of difficult and expensive work, which we would want to get done as quickly as possible because the sooner we revised the evaluation, the better we could evaluate the trajectory of change. So what does it mean to say that model revision based on the latest data should be done at "useful intervals"? It means a frequency that gives us sufficient time to reconfigure an evaluation so as to adequately assess unanticipated developments.

Of course that is not a very satisfying definition because the time needed to adjust an evaluation design is exquisitely dependent on what is being evaluated, the changes that are needed, and the structure of the existing evaluation design. My plea is that when evaluators first begin to think about budgets, resources, timelines, and methodology, they devote some time and effort to the task of planning systematic iteration between model and methodology over the lifecycle of their evaluations.

Figure 1 seems like a good idea. After all, models are needed to do powerful evaluation, and to be useful, models have to stay current. But is this a good idea? Or more correctly:

Is it good enough to be worth doing as a routine part of program evaluation?

- Are there some settings where it is more worth doing than in others? □ What is the best way to make it happen?

These are empirical questions whose answers are as yet undetermined. My hypothesis is that we would all be better off if we worked at getting answers. To get those answers we need to have a good understanding of:

Why and when program outcomes are predictable or unpredictable.

- How change happens.
- How unintended consequences can be detected.

Why and When are Program Outcomes Predictable or Unpredictable?

I find it useful to think of unintended outcomes as ranging on a continuum from: "We could have anticipated it if only we looked carefully", to "It does not matter how thorough we are, complexity dynamics make it theoretically impossible to anticipate what happened". Figure 4 Illustrates what I have in mind. (For symbolism, I used a simple tiled plane on one side, and a path of a drunkards walk on the other.)

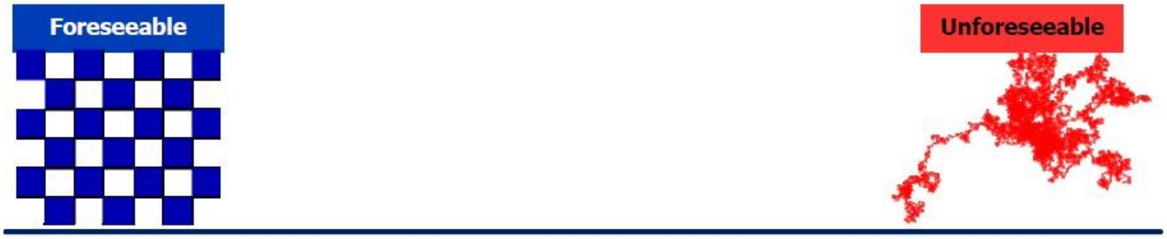

Figure 4: Depiction of Foreseeable <> Unforeseeable Continuum

The idea of such a continuum informs the role of lead time as a tactic for improving the ability of evaluation to deal with unintended consequences. At the extreme left are a variety of obvious but too seldom used tactics:

- thorough literature reviews,
- diverse input to program and evaluation design,
- asking experts who have worked with similar programs, and
- scrutiny of other programs that are in essential respects like the program being evaluated.

There is nothing new or surprising about this list. They are all good ideas, and they are all depressingly lacking in a great deal of evaluation.

Methods change as one moves from left to right. Moving to the right, the notion of "lead time" becomes ever more important because capturing knowledge will not help if there is insufficient time to make the necessary adjustments. With that movement, the importance of "agile methodology" becomes more important. At the extreme right, "agile methodology" transitions from a useful component of evaluation, to a critical one.

By "agility" I mean methods that preserve data quality and the ability to make causal inference despite changes in what needs to be measured. To illustrate, think of the "civil society" example depicted in Figure 2. In that example I suggested that validated survey scales were needed over time to assess people's beliefs in the effectiveness of their organizations. I still like that methodology, but it is fragile with respect to structure and content. With respect to structure, if baseline data collection fails, later data collection would not very useful. With respect to content, validated scales cannot be changed easily. It there is a need for even a slight shift in what needs to be known, the instrument becomes very much weaker. An alternate approach might be interviews using some kind of narrative building around people's recollection of critical incidents. I don't think the data would be as good, but it would not be awful, and it is a less breakable and more flexible design. Those are the kinds of choices that need to be made.

Of course the whole idea of the continuum is logically fallacious because it implies that it is possible to know where one is on it. If we believe (as I do) that complex behavior is abroad in the land, then unexpected change can happen at any time. And, who is to say that some level of predictability is impossible within the context of unexpected change? (In fact, this is an important notion with respect to strange attractors in the non-linear dynamics of chaos. (Strogatz, 1994), pages 324-325. Evaluators may not deal with formal strange attractors, but as metaphor, it is worth noting that such behavior does exist in nature.)

My response to these objections follows the apocryphal story of the sign in the airplane factory. "Science has proved that because of the ratio of wing surface area to body weight, the honey bee cannot fly. The bee, fortunately being ignorant of these scientific truths, goes ahead and flies anyway, thereby making a little bit of honey each day". What I have found in my work is that despite the logical flaws in the idea of the continuum, it is useful in framing evaluation methodology with respect to unintended consequences.

How Does Change Happen?

How does change happen? This question is important for two reasons.

- The answer bears on program theory. Knowing how change occurs may affect what we believe about how the program works and what consequences its actions may have. Outcomes may be different than what one expected. Or, expected outcomes may follow an unexpected trajectory. For instance, they may grow and disseminate more or less quickly than expected. Or they may grow more incrementally or discontinuously than anticipated.
- There are methodological implications. We can do a better job of evaluating unintended consequences if we have a sense of where to look.

My emphasis in this section is on the structure of change, not its domain content. To illustrate, specific reasons for untended change in programs that provide post-natal care to women will be different than reasons for unintended change in agricultural technical extension programs. Methods for dealing with these content related issues will be dealt with in the next section "How to detect unintended consequences". My intent in this section is to highlight dynamics rooted in system behavior, human cognition, group behavior, and program design, all of which cut across content domains.

In what follows I leave out a fundamental dynamic, namely, that we are working in settings where a very large number of rare events can influence outcomes, where environments are evolving and unpredictable, where we are forced to draw somewhat arbitrary boundaries around systems, and where it is uncertain as to when, how, or if activity at one level of detail affects activity at other levels of detail in ways that are *meaningful with respect to outcomes*. This is the nature of complex systems. These truths are axiomatic, and so ever present that they do not provide practical guidance as to how to design an evaluation. (They do, however, provide deep insight into why so many aid programs do not work out (Ramalingam, 2013). Please do not assume that these categories are independent. In fact, they are not. But for the purposes of explanation, it is easier to treat them separately.

Why are Unintended Consequences Likely to be Undesirable?

One further point. If there are unintended effects, they are likely to be undesirable. This opinion might just be rooted in my personality. But beyond personality, I can draw on two theoretical perspectives and the observations of others to support my dark view.

One perspective is rooted in the nature of systems and how they react to efforts at optimization. Almost every program I have ever read about or evaluated pursued either one single objective, or a set of objectives that were highly correlated with each other. By focusing all efforts and resources in such a unitary way, a program may succeed in what it is trying to do. However, the entities being affected by these programs do not have only a single need. They have multiple needs, many of which may be competing. To thrive, they need to jointly optimize multiple goals. So by moving those entities in only one of the many ways they need to change, the program may well make those entities, less, rather than better, able to thrive in the multi-dimensional world in which they live.(Jonathan A Morell, 2016).

A second theoretical perspective on the preponderance of negative consequences can be found in the work of William Easterly, who makes the point that the entire structure of aid runs on organizational relationships and funding streams that distort local conditions and local dynamics (Easterly, 2007). It is those distortions he argues, that mitigate against the productive effectiveness of change efforts.

Finally, I take comfort in the fact that my opinion is echoed in the experience of others. (Deaton, 2014) pages 312 and 317, as quoted by (Koch & Schulpen, 2017)).

> "Negative unintended consequences are pretty much guaranteed when we try. And when we fail, we continue on because our interests are now at stake – it is our aid industry, staffed largely by our professionals, and generating kudos and votes for our politicians – and because, after all, we must do something… Large-scale aid doesn't work because it cannot work, and attempts to reform it run aground on the same fundamental problems over and over again. Bridges get build, schools are opened, and drugs and vaccines save lives, but the pernicious effects are always there"

Incremental Change

Profound change can be incremental and subtle, and perhaps, not even obvious to those who are running the program. As an example, consider the civil society scenario I used above.

Example – Incremental Change

The original program was designed to work with existing organizations (Figure 2). How might the program have evolved to the one depicted in Figure 3? There may be nothing obvious about how it happened. Perhaps a few engaged citizens knew of the program's activities, and began asking program staff for help in setting up a new organization. Such a change in program services might begin with nothing more than a few staff members using their slack time to provide a little technical assistance. The beginnings of the change may never even register, and the evolution toward helping new organizations might be so subtle as to be accepted and invisible to program staff as a change worth noting.

Discontinuous Change

The previous section dealt with incremental change. It is also possible that the outcomes of a program will maintain themselves in a stable and expected pattern, and then change in sudden and dramatic ways.

Network Effects

Rapid change after a period of quiescence is possible any time the outcomes of a program affect the richness of relationships among entities. Networks are like that. Growing connections seem to have little effect on the network structure until a state change takes place that quickly brings the network together. Figure 5 is an illustration of this effect (Wilensky, 2005). (The reference will point you to a Web resource that will let you run dynamic simulations of this process.)

As long as one knows that program outcomes may have networking effects, this kind of change in outcome is well anticipated. After all, there is a great deal of research on network effects, and we know the program is designed to enrich networks. That is the case in the civil society example that I have been using. But what if the network building phenomenon were not part of the planned outcome of the program, but was taking place anyway? Here is an example.

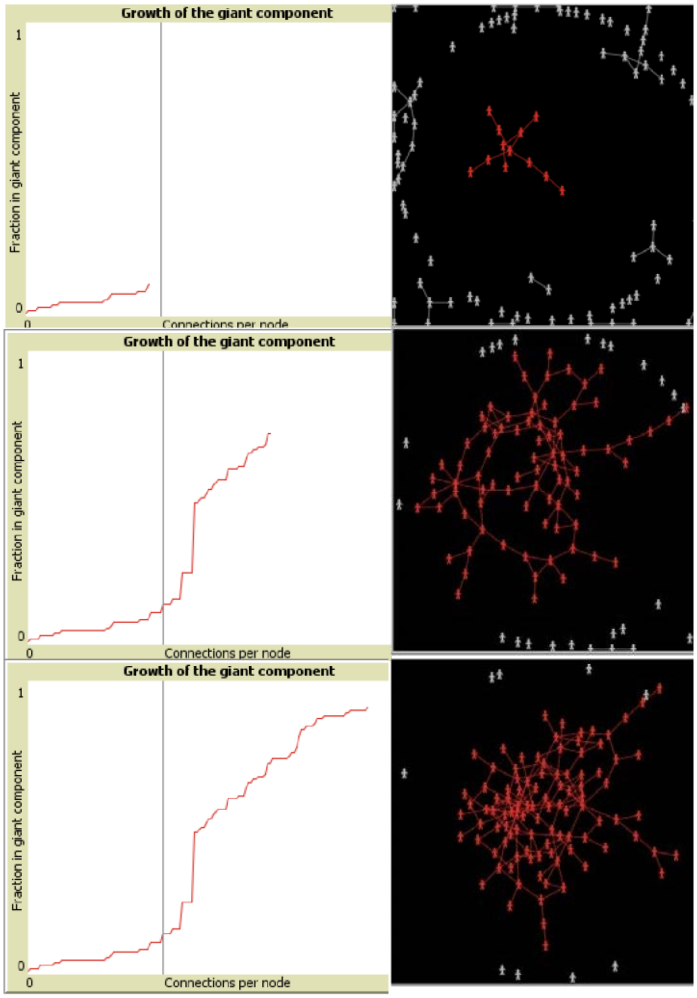

Figure 5: Network Growth

<u>Example: Unanticipated Network Outcomes</u>

Think of a program that is designed to provide support to small and medium sized enterprises (SME) in a particular region of a developing country. The program services are straightforward business type activities, e.g. process control, marketing, supply chain relationships, inventory management, accounting systems, business planning, and so on. The outcome chain is also straightforward. The immediate outcome is higher profit. The follow-on outcomes include the obvious consequences of profit, e.g. housing quality, children's education, access to health care, and so on. Community effects are also hypothesized as a result of the newly profitable companies spreading money throughout their communities.

Parallel to these activities, however, network effects might be building. People who run businesses know each other personally, and also interact in meetings of groups analogous to chambers of commerce. As each company becomes more profitable, and thus better able to explore new opportunities, business arrangements among companies may develop. Also, already existing dependencies among the companies may become stronger. If that were the case, there may suddenly appear changes as a function of a state change in a network that nobody even knew was forming. I am not claiming that this networking effect would occur, only that it could. Most important for our purposes, the network effect was never present in the original set of expected outcomes, and hence omitted from whatever model was built to guide the evaluation.

Feedback Loops

Evaluators have long recognized that feedback loops play an important part in modeling (Dyehouse, Bennett, Harbor, Childress, & Dark, 2009). Also, it is well known that even simple feedback loops can cause non-linear behavior (Wikipedia, 2016c). So why include this subject in a discussion of unintended consequences? After all, if we suspected such behavior, we would address it in our methodology. In any case, in our world, nobody would knowingly give those feedback dynamics free reign. We manage our programs. If undesirable behavior were taking place, we would do something about it.

What makes it relevant for a discussion of unintended consequences is that we do not usually attend to feedback loops among outcomes. We do not usually recognize the possibility that an outcome may affect what appears to be a logically prior outcome. Such a dynamic, however, can indeed occur. As an example, consider the (hypothetical) example represented by Figure 6.

Example: Feedback Loops Among Outcomes

Figure 6 depicts the outcomes of a program designed to increase entrepreneurial behavior. The depiction on top is the basic logic of the outcomes that the program designers and evaluators worked out. Entrepreneurial behavior leads to new business activity, which leads to a change in the standard of living. It is possible that the evaluation would include the feedback loop, but in my experience, it probably would not. (By the way, we should not assume a directionality that leads to higher standards of living. It is quite possible that so many of the new businesses would fail that the standard of living would go down, as would the level of entrepreneurial behavior.)

The bottom view depicts a set of feedback loops that are by no means unreasonable. There, the average standard of living does indeed increase, but at the same time, so too does inequality. Given the nature of capitalism, this is not an unreasonable conjecture. One obvious problem is that the evaluators failed to account for this outcome, thus greatly increasing the likelihood of encountering unintended consequences. But from the point of view of the present discussion, the notable feature of the model is the existence of a new set of reinforcing feedback loops. Those could have truly dramatic effects on the behavior of the entire system, and yet as a mechanism driving those consequences, they would be completely invisible.

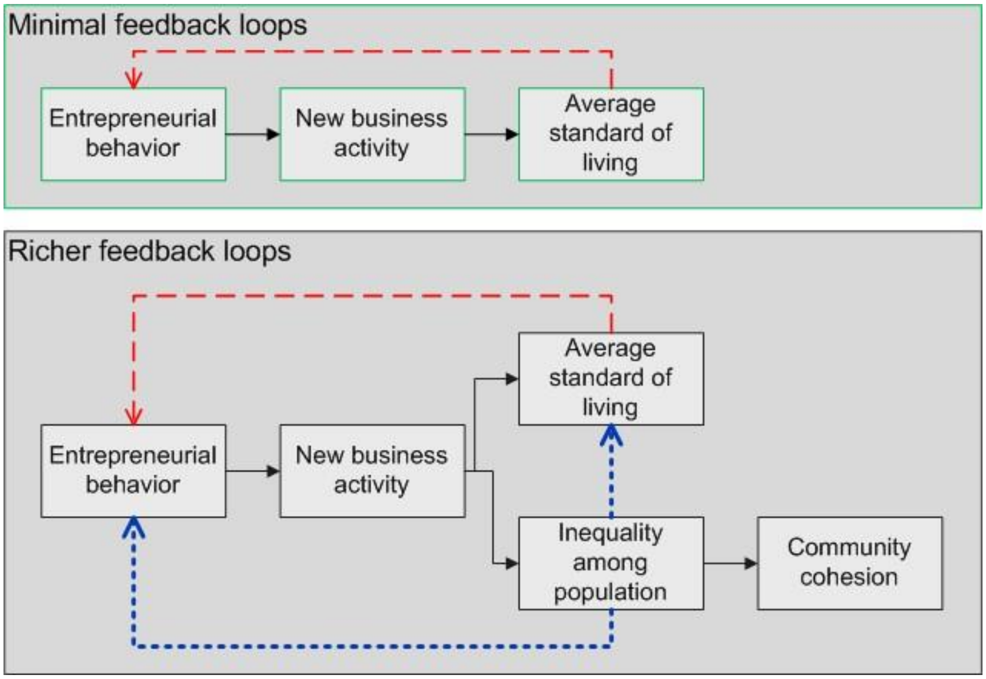

Figure 6: Accounting for Feedback Loops

Social and Psychological Sources

The previous items in this section all dealt with structural and system-based reasons why change happens. It is also worth considering more human sources. It is trite to assert that we are all human, and that humans make mistakes. But it is worth considering what it is about us that causes us to generate error when it comes to the task of identifying program outcomes.

Our individual reasoning processes are suffused with cognitive biases that keep us from making rational judgements. To name just a few: "Anchoring" has us attaching undue weight to information that we come upon first. "Base rate" leads us to ignore base rate information when focusing on specific events. The "Illusion of control" has us overestimating our influence over events. "Optimism" leads us to overestimate the likelihood of desirable outcomes (Klocke, 2007; Wikipedia, 2016b). We are terrible at distinguishing signal from noise (Silver, 2012). All of these, and many others I have not listed, drive our chronic inability to establish accurate project schedules, or to be conscious of what all the consequences of our actions will be. (See a discussion of the "planning fallacy", chapter 23 in (Kahneman, 2011). We carry these individual cognitive biases into our behavior in groups, thus adding another layer of distortion to our reasoning. Our activities in groups are rife with dynamics such as talkative and/or high prestige people driving deliberations, and conflict aversion tamping down minority or divergent opinion.

Program Design

Unintended consequences can arise from the intersection of system behavior and the social/economic/political drivers of program design. (In this section I am building on the previous discussion of why unintended consequences tend to be undesirable, and drawing from the more elaborate discussions in (Jonathan A. Morell, 2016a, 2016c).

Pick your favorite program and observe the outcomes in its logic model. If your experience is like mine, you will find that the result of the logic model building exercise is a set of outcomes, some of which will be antecedent to the end of an outcome chain. Also, there may be more than one outcome in the terminal position. But one thing is almost certain. All of the outcomes will be highly correlated. Whatever the desired directions of change, if one outcome gets better as a result of program action, so too will the others. I understand why programs are like this. Programs are implemented by people who are embedded in organizational settings that are characterized by well-defined boundaries. Within those boundaries:

- Reward systems differ.
- Political and mission priorities differ
- Funds are dedicated for specific purposes.
- Disciplinary paradigms are difficult to bridge.
- Organizational boundaries are hard to transcend.
- Coordination becomes more difficult as scope widens.
- Coalitions in favor of a program grow weaker as design compromises pile up.
- If you want an in-depth explanation, see the video at: (Jonathan A. Morell, 2016b).

On the one hand, any exercise that gets program designers to appreciate their constraints, to assure alignment between action and mission, and to identify outcomes, is a very good thing. But there is a dark side that touches both outcomes, and the workings of the programs that are designed to effect those outcome.

Outcomes

Pick anything that might be the target of an intervention to make things better – people, communities, school systems, county governments, neighborhoods, universities, wetlands, road systems, farms, export policies, democracy promotion programs, or anything else that piques your interest. I think of these as organisms working at climbing their fitness landscapes, but I know that many people do not like that way of thinking. So here I'll just call them "boxes". Whichever box you pick, I bet that a close look will show that for that box to thrive over the long term, it will need to meet a variety of needs. Whatever these needs may be, I'm pretty sure that:

- Their time-frames will vary.
- They will not all be equally important.
- They will not have to be achieved at the same level success.
- Some will be more robust in the face of challenge than others.
- Some may spring into existence where they did not exist before.
- The same need may carry different salience under different conditions.
- Some may be latent and only become manifest under particular conditions.
- Some will become important because of workings inside the box, and some because the box needs to negotiate its relationship with its surrounds.
- But whatever their specifics, there are sure to be more than one.

While there may be times when a single acute need takes precedence over all the others, eventually more than one will need to be pursued. So what will happen if a program comes along that pushes a box to meet only one of its goals to the exclusion of the others? I do not know, but whatever it is, it will not be good. Eventually the box will have to do something to thrive. (I am speaking as if the box has volition, which in many cases it will not. But for explanatory purposes, speaking in terms of volition has its advantages.) The box has two choices. It may:

- Ignore the objective the program is promoting, or
- pervert the pursuit of the program's goal to serve other needs.

How many uncorrelated outcomes are needed to produce a "healthy" box? I have no idea, but I am sure that:

- One alone is problematic.
- Too many will strain the resource and capabilities of the program.
- At some point multiple goals will probably become fairly correlated anyway, so there is no point in chasing too many.

Programs

The most optimistic possible scenario is that a program could, if it wanted to, implement initiatives that would jointly optimize multiple goals. I believe though, that once programs commit to a goal, it becomes very difficult for them to adapt to new needs. In essence, the choice to help boxes meet particular objectives drives a course of program development that makes it difficult for a program to do anything else. And as they keep pushing boxes toward a single goal, the programs generate more of the unintended consequences that they would prefer to avoid. This can be a very destructive feedback loop that explains a lot about the source of undesirable consequences.

Why does this vicious cycle exist? Because programs can be thought of as systems, and systems do not thrive over the long term if all of their resources, and all of their actions, and their entire structure, are dedicated to only a single outcome. Those kinds of systems:

- Are rigid, and break easily.
- Have limited capacity to monitor their environments.
- Have minimal internal capacity to adopt to new circumstances.
- Ignore the reality that even if clients have a single overriding need over the short term, they are sure to have varying needs over the long term.

The last is most important. Pursuit of a single goal, or multiple goals that are highly related, forces systems to follow one or more paths:

- Motivate boxes to opt out of the program.
- Ignore other needs that boxes are sure to have.
- Find ways to pervert themselves in the service of those other needs.
- Oscillate among some of the above options.

Whichever path is pursued; the result will not be good.

Why is it Hard to Discern Change?

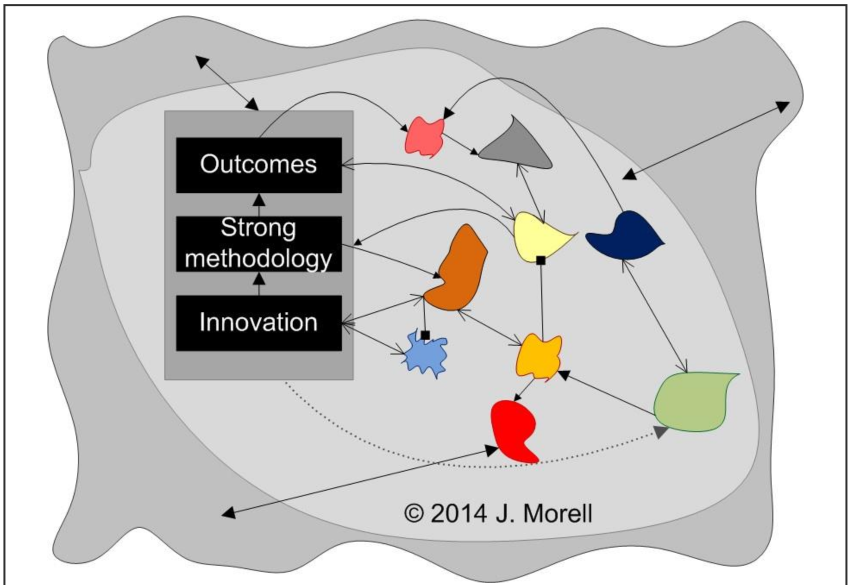

Figure 7: "Simple" Programs Embedded in Rich Settings

If it were easy to detect change, we would do it. The reason we have trouble is that all of our evaluation designs assume a false simplicity. They must, because the programs they evaluate are also built on models, and as we know, to be useful, models have to be simplifications of reality.

Figure 7 conveys a sense of the situation. We work with what appears to be a well-defined methodology to evaluate a well-defined program. Our efforts, however, are embedded in a setting that is full of other elements that may be influencing our program, and that are rich with network effects, feedback loops, and all manner of transitory fluctuations that might align to drive change.

Under these conditions, is it so hard to believe that we miss important change that starts with the transient alignment of fluctuating conditions, or that we fail to see networks forming or feedback loops in operation? It is hard to see these phenomena, and that is why we need a rigorous process to detect them.

How to Detect Unintended Consequences

This paper began with an explanation of why models are so important in evaluation, and why the utility of models is limited if they are not continually updated based on empirical evaluation data. The second section of this paper discussed principles of system behavior, human cognition, group dynamics, and program design that can explain how change happens. That knowledge is useful for determining what patterns of change our models need to accommodate.

While I did not touch on it specifically, I have made the assumption that a sound evaluation is already in place, i.e. one that is attentive to expected outcomes, to internal program behavior, and to environmental conditions, and with attention to causal inference by means of a nice mix of structured and less structured methodologies.

Now the question presents itself. What would be an effective and efficient way to conduct the interplay between new data availability and model revision? The answer depends on making wise choices with respect to:

□ The people who are included in the activity, and

□ the work those people are asked to do.

Who to Include

Data interpretation and model development are most insightful when they represent diverse perspectives. For the purposes of discerning unintended consequences through an interplay of data and model, five perspectives are useful. In what follows I use the term "people" for reasons of language flow, but what I really mean are "points of view", or "a group's interests". "People" is a useful term because they are the most likely sources of information, and because there are particular ways to extract information from them.

1. "Close observers".

   These are people close to the program who have with rich opportunity to observe the program's operations and outcomes. Examples of such people are program staff, collaborators in other organizations, and so on.

   A special case of "close observers" are people who are likely to experience what a program does, either directly, or indirectly.

2. Representatives of systems that are likely to be affected by a program's outcomes. As an example, educational systems may see changes in their students as a result of a nutritional or health promotion effort.

3. Evaluators not connected to the program, but who have experience with similar programs in other settings.

   What is needed from this group is knowledge of what has happened with other similar programs. If enough previous programs have resulted in particular outcomes, it is a safe bet that those outcomes will manifest again, regardless of anything that seems to argue for the program in question being unique or special.

4. People with no direct connection to the program, but who, because of their backgrounds, may have opinions that are worth considering.

   As an example, imagine a program whose purpose was to provide technical assistance in developing countries on matters related to agriculture. In such a situation I could imagine the value of including people who have worked in that country on topics such as small business development, export promotion, and nutrition. These would be points of view that were informed by knowledge of the local setting, expertise in providing development assistance, sensitivity to various remote effects that may emanate from agricultural technical assistance, and prior exposure to evaluation exercises.

5. The categories above all represent points of view with respect to a program's activities. A different take on diverse points of view touches on the intellectual foundations of the program.

   Let us go back to the civil society example. I have no doubt that if the evaluation data were perused by a political scientist, an economist, and a sociologist, there would be rather different opinions about the value of the observed outcomes, the secondary effects of those outcomes, and the model needed to capture the impact of the program on its environment.

The sampling scheme presented above has the strong advantage of including a wide range of background and expertise. It also has the strong disadvantage of including a wide range of background and expertise. There needs to be a way of managing the currents of opinion, bias, and personality that are sure to swirl. The objective is to elicit expert opinion that is founded on each participant's expertise, and to identify those areas of agreement and disagreement among them. With that knowledge in hand, wise choices can be made with respect to revising the model and the data collection plan. There is no need to achieve consensus, which even if it could be reached, would likely paper over meaningful disagreement. Nor is there a need to do what the data interpretation advisors advise. There is only a need to make a good faith effort to take the input seriously.

How to Include Them

The Delphi technique is designed precisely for situations where conflict, bias, and group dynamics may interfere with dispassionate analysis. In it, individuals are queried individually for their opinions. Areas of agreement and disagreement are then identified and another round of individual communication takes place, each time asking people to elaborate on the reasons for the disagreement. In this way individuals know the opinions of others and can respond to them without the fuss and expense of face to face interaction (Linstone & Turnoff, 1975; Wikipedia, 2016a). In order to make this process work for managing the interplay of model building and data interpretation, I suggest the following.

- As a foundation, make sure that all involved are familiar with the program, the model, and the current evaluation findings.

- Ask each member of the group to suggest further analyses that they would like to see. Of course there is no guarantee that the evaluators will be able to accommodate those requests, but good faith effort must be made.

- Ask each member of the group what a combination of the data and his or her experience reveals about the outcomes that have been observed, and outcomes that are likely to take place.

    - With respect to these outcomes, ask people about their level of certainty, and why they think those outcomes will occur. This latter question is critical for any revisions of the model that may be developed.
    - As these queries are made, the panel's attention needs to be drawn to the sources and types of change discussed in the previous section, as a way of helping them think creatively about what consequences the program might have, and the mechanisms underlying the appearance of those consequences.
    - While it may suffice to simply pose questions to the panel, it is worth considering a variety of structured ways of interacting with them. Scenario based planning envisions multiple futures, and then posits development paths to get there (Godet, 2000; O'Brien, 2003). Backcasting assumes a future state and then asks "How did we get here?" (Drenborg, 1996). Assumption Based Planning tries to identify different types of elements needed to achieve a goal: 1) critical assumptions that must be met, 2) load bearing assumptions that are susceptible to breakage, 3) signposts that are indicators of potential problems, 4) shaping actions that can support uncertain assumptions, and 5) hedging activities to prepare for the possible failure of critical assumptions (Dewar, 2002). Also, in general, when it

comes to scouting for surprises, there is much to be learned from the extensive knowledge residing in the field of project management (Kerzner, 2013; Larson & Gray, 2011).

- Distribute a summary of the findings to the group, along with a request for reactions.
- Conduct further rounds until there is a sense that new knowledge will no longer emerge.
- Once the process is over, distribute the new model to all involved, along with an explanation of the revised methodology and the outcomes that are going to be measured.
- At the next round of data collection, repeat the process.

There is no formula to determine precisely who should be involved in these exercises, or how many people should be included. As a rule of thumb I offer the following. First, more than one member of each of the five participant groups is needed. Second, there is no obligation to use all of them at each round of analysis and model development. Also, there is no requirement that the same people be included each time the model is revisited. There is a good reason to use the same people because there is advantage in working with people who are familiar with the process. On the other hand, it is entirely possible that a program will evolve over time, or new sets of outcomes will be manifest that would benefit from expertise not represented in the original group.

Funders and Funding Requirements

The activity described above assumes that funders of evaluation would welcome a determined effort to evaluate unintended consequences. It is by no means obvious that this is so. In fact, my experience is that it is not. One hindrance is cost. But money notwithstanding, allowing these efforts requires a belief that unintended consequences will occur and that they are worth evaluating. If I am in the business of funding programs that improve girls' access to education, do I care about whether the program will also affect people's health status? I do as a human being interested in the social condition. But as a policy maker accountable to a particular set of government functions and funding streams? Not so much. Finally, how much motivation would I have to invest in the evaluation of unintended consequences if I had an inkling that many such consequences would be undesirable? Thus before all those experts can be engaged in the model building, some effort may be needed in the political and policy arenas.

A (hopefully realistic) Determination of Level of Effort

I have no illusion that what I am suggesting is cheap, easy or comfortable. There are considerable challenges in developing and revising models. Nor is it a trivial matter to recruit participants, organize their activity, redesign evaluation, or implement new designs. Also, doing evaluation as I am suggesting would require deep changes in the culture of the evaluation community, beliefs within programs about what constitutes good evaluation, and the expectations of funding agencies.

On the other hand, consider the cost that would accrue to efforts to further international cooperation if a process like this were not included in evaluation. This entire conference is taking place because of a belief that there are unintended consequences, and that ignorance of those consequences drives ineffective or counterproductive action.

So, how much effort to put into revising the model at multiple times over the evaluation’s life cycle? The answer depends on estimating the likelihood that unintended effects will occur. I hope that what I have said earlier about the behavior of complex systems makes it clear that answering this question is impossible. But as with the foreseeable □□ unforeseeable continuum, the honey bee metaphor is still valid. Attempting an estimate is still a good idea. What to look for in making that estimate? I find it useful to ponder the answers to five questions:

1. How rich and tight are the linkages among major elements of the program? Rich, tight linkages make for unpredictability in system behavior. We may not care whether the program operates as planned, but we do care if it behaves in ways that will elicit unintended outcomes.

2. What is the “size” of the program relative to the boundaries of the system in which it lives? Small programs have less of a chance to behave in ways that will perturb the systems in which they reside. And when systems are perturbed, they behave in unexpected ways.

3. Where is the program in its life cycle? Programs in the start-up phase may change rapidly. As with item #1, the change itself is not a problem. What is problematic is change that affects outcomes.

4. How stable will the environment be between the time we implement the innovation and the time we expect outcomes?

5. How robust is the innovation across time and place? There are two aspects to robustness, one internal to the program and one external. The internal aspect is the extent to which the program has latitude in its activities. Or put another way, how faithful does the program have to be to a defined set of criteria (Abry, Hulleman, & RimmKaufman, 2015; Zvoch, 2012)? The external aspect is the range of diverse settings in which the program can be trusted to manifest a known set of outcomes.

If you conclude that unintended effects are likely enough to be worth attention, then two questions become relevant:

1. How much attention needs to be paid to increasing the lead time between when such changes are suspected and when they occur, and

2. What design tradeoffs are acceptable to make sure the evaluation design is adequately agile?

These questions are important because design change can weaken methodological rigor with respect to evaluating intended consequences, which, presumably, have not gone away. Consider the possible ramifications of changing an evaluation design once it has been implemented. 1) Different intellectual resources that are not readily available may be needed. 2) There may be budget implications. 3) Stakeholders’ expectations may be affected. 4)

Contractual changes may be needed. 5) The mental health of the evaluators will be challenged. 6) Any change in the design will require a tradeoff with respect to the logical structure of the design. Thus, tradeoffs between the original intent and the new evaluation purposes will be necessary. (For a process flow diagram of the consequences of these tradeoffs and the opportunity costs incurred in making them, see chapter 7 in (Jonathan A. Morell, 2010)).

If changing an evaluation design has costs, how might evaluators and donors decide if the costs should be incurred? The answer depends on how capable you guess that the existing design will be in addressing potential unintended consequences. Making this estimate is one more of the challenging tasks that I have suggested in this paper, but which are still worth attempting. The estimate will depend on answers to three questions.

1. Will I be able to use existing data, with the data being current starting from the time when I first realize I will need it?

2. How important will it be to make comparisons by drawing on historical data, i.e. data that predates the time when the need for that data first became apparent.

3. How important will it be to draw data from some kind of comparison group?

Or to phrase the questions in a different way. How confident could I be in my evaluation findings if:

- The methodology consisted of a "posttest only", no control group design.
- Archival data consisted only of easily records.
- Interview data came from unstructured, or semi-structured questions.

This is an inherently agile design. If it will suffice, use it.

Research Agenda

The title of this paper includes the phrase: "From Firefighting to Systematic Action". The premise is that there are ways to:

- anticipate unanticipated consequences, and to
- do a better job of evaluating them when they truly surprise us.

The overarching hypothesis posed is that the logic in Figure 1 will lead to evaluation practice that is better able provide decision makers with good information about the positive and negative unintended consequences of their actions. Along the way many research questions have surfaced. Below I offer a list of research I that think is important, along with a sketch of how the questions might be addressed.

Research Question	**Sketch of Research**
Is the basic premise of the argument correct?	
1- Can evaluators do a better job of addressing unintended consequences if they continually iterate between model building and data collection over the course of an evaluation's life cycle?	Methodology / research logic ▪ Multiple comparative case studies based on existing or soon to be implemented evaluations. ▪ Follow-up for as much time as possible with respect to the use of use of evaluation results. Data ▪ Interviews with evaluators. ▪ Interviews with funders. □ Document review. Research implementation ▪ Practical because logic models are required anyway, so the only innovation is a change in accepted practice. ▪ Requires a central core of researchers and a mechanism to coordinate with the ongoing evaluations.
2- Are there some conditions for which it is particularly worth the effort?	Methodology / research logic ▪ Literature review and interviews with experts to identify likely important factors, e.g. novelty of innovation and/or setting, history of similar programs, relative emphasis on implementation fidelity vs. a developmental approach. ▪ Stratify on domains to control for context-specific factors, e.g. AIDS prevention, development of civil society, primary schooling for girls. ▪ Choose ongoing evaluations if necessary, but best to deliberately recruit in advance of implementation. ▪ Do analysis based on hypothesized important factors, but also use a grounded theory approach to compare dramatic cases where use of modeling was and was not of value. Data ▪ As in question #1. Research implementation □ As in question #1.

What are effective methods of determining when model and methodology should interact?	
Calendar driven Use whatever data are available at fixed times in the evaluation's life cycle. Timing to events in the evaluation e.g. after a particular batch of data are collected, or when an interim report is due. Timing to phases or important events in the project's life cycle For instance, these might be when new services are implemented or outreach to new groups begins. Timing to what are deemed to be unplanned critical events. One example may unexpected policy changes affecting a program's operations, or a drop off (or increase) in demand for a program's services.	Methodology / research logic ▪ Multiple case studies as in question #1. ▪ Stratify on domains, as in question #2. ▪ Stratify on timeframe of evaluation, e.g. six-month life cycle versus a three-year life cycle may benefit from different schedules. □ Post-hoc comparisons as in question #2. Data ▪ As in question #1. Research implementation ▪ Necessary to work closely with evaluation teams to make sure they understand the circumstances and timing of their data/model interactions. ▪ Requires a central core of researchers and a mechanism to coordinate with the ongoing evaluations.
What are effective types of diversity with respect to input on models?	

▪ Evaluators with experience in similar programs □ Funders. ▪ Service recipients. ▪ Domain experts.	Methodology / research logic ▪ Primary comparison is between "insider" input (funders, service recipients), and "outsider" input (domain experts, evaluators with experience in similar programs). ▪ Case study and stratification as above. ▪ Two stage interaction – Delphi first to minimize group process bias, followed by face to face meetings. Data ▪ Observation, analysis of deliberations ▪ Observation of relationships between model revision and evaluation data collection □ Observation of usage of evaluation results. Research implementation ▪ Primary collaboration through digital media.
How do evaluators respond to unintended consequences now? (There is a need to expand small amount of existing research.)	
▪ Methodology ▪ Interaction with program staff □ Interaction with funders	Methodology / research logic ▪ Literature search for dramatic cases, post hoc analysis. (Advantage = short time frame, large "effect size". Disadvantage = cannot do in-depth analysis as events unfold in real time.) ▪ Augment post-hoc sample with longitudinal case observation. Data ▪ Interviews with evaluators and funders ▪ Document review Research implementation ▪ Primary difficulty is interviewing people about sensitive issues. Unintended consequences often means that either the program did not work out as planned, or the evaluation did not, or both. In all cases, the topic is painful to all involved.

Policy Recommendations

This paper has attempted to explain why unintended consequences of program action occur, and to argue that evaluators could better understand those consequences if they adopted a practice of combining modeling and data collection over the course of an evaluation's life cycle. There are messages for both program design and funding, and for research to improve the ability of evaluation to deal with unexpected consequences.

Program Design and Funding

The use of logic models, and a related emphasis on program theory, are deeply accepted aspects of funding for programs involving international cooperation. Those models, however, do not reflect many of the dynamics by which programs actually operate, as discussed in the sections "Why and When are Program Outcomes Predictable and Unpredictable?", "How Does Change Happen?" and "Why is it Hard to Discern Change?". Those sections point to dynamics of program outcome that are not captured in traditional logic models or theories of change. And because they are not captured in those models, they are not expressly represented in evaluation methodologies or data collection plans. Program designers and funders would do well to work with their evaluators to include those kinds of behaviors. Doing so, however, would require a change in the mindset, culture, and accepted practice of funder/evaluator interactions.

Better Evaluation Through Data/Model Interaction

This paper has suggested a research agenda to improve evaluation's ability to deal with unintended program outcomes. Organizations that have an interest in better programming would do well to support that research. While some of the needed resources would have to be added to support for routine evaluation work, much of the research could be integrated into ongoing evaluation activities. In that sense, much of the needed work is organizational and procedural, rather than resource-consuming.

Thus in both policy domains – "program design", and "better evaluation" – valuable progress is possible by changing the nature of funder/evaluator interactions. A few small-scale, pilot test efforts in this direction could serve as a foundation for more extensive activity.

References

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. E. (2015). Using Indices of Fidelity to Intervention Core Components to Identify Program Active Ingredients *American Journal of Evaluation, 36*(3), 320 338.

Deaton, A. (2014). *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.

Dewar, J. A. (2002). *Assumption Based Planning: A Tool for Reducing Avoidable Surprise*. New York: Cambridge University Press.

Drenborg, K. H. (1996). The Essence of Backcasting. *Futures, 28*(9), 813 -- 828.

Dyehouse, M., Bennett, D., Harbor, J., Childress, A., & Dark, M. (2009). A comparison of linear and systems thinking approaches for program evaluation illustrated using the Indiana Interdisciplinary GK-12. *Evaluation and program planning, 32*, 187–196.

Easterly, W. (2007). *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and o Little Good*. New York: Oxford.

Godet, M. (2000). The art of scenarios & strategic planning: Tools & pitfalls. *Technological Forecasting & Social Change, 65*, 3 - 22.

Kahneman, D. (2011). The Outside View *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kerzner, H. (2013). *Project Management: A Systems Approach to Planning, Scheduling, and Controlling, 11th ed.* NY, NY: Wiley.

Klocke, U. (2007). How to Improve Decision Making in Small Groups: Effects of Dissent and Training Interventions. *Small Group Research, 38*(3), 437-468. doi:10.1177/1046496407301974

Koch, D.-J., & Schulpen, L. (2017). *Unintended effects of international cooperation: A preliminary literature review*. Paper presented at the Unintended Effects of International Cooperation, The Hague, The Netherlands.

Larson, E., & Gray, C. (2011). *Project Management: The Managerial Process, 6th ed.* NY, NY: Mcgraw Hill.

Linstone, H. A. e., & Turnoff, M. e. (1975). *Delphi Method: Techniques and Applications*. Reading, MA: Addison Wesley (Pearson).

Mertens, D. M., Bazeley, P., Bowleg, L., Fielding, N., Maxwell, J., Molina-Azorin, J. F., & Niglas, K. (2016). Expanding Thinking Through a Kaleidoscopic Look Into the Future: Implications of the Mixed Methods International Research Association’s Task Force Report on the Future of Mixed Methods. *Journal of Mixed Methods Research, 10*(3), 221 - 227.

Morell, J. A. (Producer). (1979). Evaluation as Social Technology Chapter 5 in. *Program Evaluation in Social Research*. Retrieved from https://evaluationuncertainty.com/evaluation-as-socialtechnology/

Morell, J. A. (2005). Why are there unintended consequences of program action, and What Are the Implications for Doing Evaluation? *American Journal of Evaluation, 26*(4), 444 - 463.

Morell, J. A. (2010). *Evaluation in the Face of Uncertainty: Anticipating Surprise and Responding to the Inevitable*. New York: Guilford.

Morell, J. A. (2016a). Joint Optimization of Uncorrelated Outcomes as a Method for Minimizing Undesirable Consequences of Program Action. Retrieved from https://evaluationuncertainty.com/2016/12/19/joint-optimization-of-uncorrelated-outcomes-as-amethod-for-minimizing-undesirable-consequences-of-program-action/

Morell, J. A. (2016b). Part 2 -- Drawing on Complexity to do Hands-on Evaluation: Why Use Complexity When Programs Do Not? . Retrieved from https://www.youtube.com/watch?v=2feakJJlY3U Morell, J. A. (2016). A simple recipe for improving the odds of sustainability: A systems perspective. Retrieved from https://evaluationuncertainty.com/2016/06/12/a-simple-recipe-for-improving-theodds-of-sustainability-a-systems-perspective/

Morell, J. A. (2016c). A simple recipe for improving the odds of sustainability: A systems perspective. Retrieved from https://evaluationuncertainty.com/2016/06/12/a-simple-recipe-for-improving-theodds-of-sustainability-a-systems-perspective/

Morell, J. A., et. al. (Producer). (2016a). Part 2-- Modeling, Agents, and Complexity: New Tools, New Theories for Program Evaluation. Retrieved from https://youtu.be/41bpexkGURI

Morell, J. A., et. al. (Producer). (2016b). Part 3-- Modeling, Agents, and Complexity: New Tools, New Theories for Program Evaluation. Retrieved from https://youtu.be/z0Tk2jIfsF4

Morell, J. A., et.al (Producer). (2016). Part 1-- Modeling, Agents, and Complexity: New Tools, New Theories for Program Evaluation. Retrieved from https://youtu.be/bNuqDjD0ZSw

Morell, J. A., Hilscher, R., Magura, S., & Ford, J. (2010). Integrating Evaluation and Agent-Based Modeling: Rationale and an Example for Adopting Evidence-Based Practices *Journal of Multidisciplinary Evaluation*, *6*(14), 35 -- 37.

O'Brien, F. A. (2003). Scenario planning – lessons for practice from teaching & learning. , . *European Journal of Operational Research*, *152*, 709 - 722.

Orrell, D. (2007). *The Future of Everything: The Science of Prediction*. New York: Thunder's Mouth Press.

Parunak, H. V. D., & Morell, J. A. (2014). *Emergent Consequences: Unexpected Behaviors in a Simple Model to Support Innovation Adoption, Planning, and Evaluation.* Paper presented at the Social Computing,Behavioral-Cultural Modeling, and Prediction 7th International Conference, SBP 2014 , April 1-4, 2014, Washington, DC, USA.

Ramalingam, B. (2013). *Aid on the Edge of Chaos: Rethinking International Cooperation in a Complex World*: Oxford.

Silver, N. (2012). *The Signal and the Noise*. New York: Penguin Press.

Strogatz, S. (1994). *Nonlinear Dynamics and Chaos*. Reading MA: Addison Wesley (Pearson).

Weisberg, H. I. (2014). *Willful Ignorance: The Mismeasure of Uncertainty*. New York: Wiley.

Wikipedia. (2016a). Delphi Method. Retrieved from https://en.wikipedia.org/wiki/Delphi_method Wikipedia. (2016b). List of cognitive biases. Retrieved from https://en.wikipedia.org/wiki/List_of_cognitive_biases

Wikipedia. (2016c). Nonlinear control. Retrieved from https://en.wikipedia.org/wiki/Nonlinear_control

Wilensky, U. (2005). NetLogo Giant Component Model. *NetLogo Models Library*. Evanston IL: Center for Connected Learning and Computer-Based Modeling, Northwestern Univeristy.

Zvoch, K. (2012). How Does Fidelity of Implementation Matter? Using Multilevel Models to Detect Relationships Between Participant Outcomes and the Delivery and Receipt of Treatment *American Journal of Evaluation*, *33*(4), 547-565.