Evaluating the future

Rick Davies, 2020 06 16

"The future is already here – it's just not evenly distributed"
William Gibson, The Economist, December 4, 2003

Background: This paper was written as part of my preparations for a podcast of the above title, following a request from the EU's Evaluation Support Service. The overall theme of the podcast series was '*How can evaluation practice be adapted in the face of the global pandemic?*'

In the recorded podcast which you can find here, I began by making these points in summary:

1. My recommendations are orientated not specifically to the pandemic, but to the wider issue of how we respond to uncertain and unpredictable events. The pandemic and its various manifestations, as seen in and by different countries, is an exemplar of these kinds of events.
2. My recommendations are of two kinds. The first is that we need a lot more practice in thinking about multiple alternative futures, rather than one desirable future. The second is that we will also need to think a lot more about how to evaluate those alternative futures. This is a kind of 'evaluative thinking 'which up to now does not seem to have received much attention.

The rest of this paper is about these two recommendations, but particularly about how we evaluate alternative futures.

Today I am going to talk about evaluating the future.

To begin with I am going to introduce the topic of **metacognition.** Metacognition is thinking about thinking. Your immediate reaction might that be this sounds very unworldly, …"*So how is this going to be of any use to me?".* But you might be surprised to know that metacognition is quite a lively topic, with real-world applications in the field of primary and secondary school teaching. This is because metacognition is not only about being aware of how we think about things but also about being able to *regulate* how we think about things. A practical example of meta-cognition is when a student thinks about planning an essay, then assess their progress in writing the essay. And how a teacher helps a student reflect on how they do this. Another example of this which you might be familiar with is mindfulness meditation, where people attend to the presence of arising thoughts, but consciously decide not to engage with them

In this podcast I going to talk about another area of metacognition, not usually addressed within school settings. This is about how we think about the future. Not only the content of those thoughts, i.e. what we think will happen, but also how we assess those possibilities. To me this is an important but neglected facet of what is called these days "**evaluative thinking**". Evaluative thinking is an aspect of evaluation practice which we would like to see everyone engage with and get better at.

Most evaluations are backwards-looking, looking at what was intended, what was delivered and what the outcomes were. However, there are some elements of looking into the future. There might be an ex-ante-evaluation of the project design. There might be an evaluability assessment. And the recommendations of an evaluation will be future orientated. But I think it would be fair to say that, in most evaluations, the scope of this thinking about the future is fairly bounded.

Another area where evaluators and program designers must pay some attention to the future is in project design itself, particularly the articulation of **Theories of Change**. This is something I have written about quite a lot over the years. Most recently in a paper commissioned by CEDIL, called Representing theories of change: technical challenges with evaluation consequences. In that paper I describe a range of methods ranging from simple chain models of how activities lead to outcomes, to more complex static network models, and on to dynamic models where the effects of an intervention can be simulated. One of the merits of dynamic models is the quick discovery of unintended consequences of interventions that can arise because of the many interconnections between events in the model. Another salutary lesson is how sensitive the status of the dynamically generated outcomes can be to the settings built into the model. As the widely repeated quote (of disputed origins) says… '*making predictions can be difficult, especially about the future* '

The **difficulty of prediction** has been brought home to us with force this year, as we have watched the global spread of the coronavirus and the devastation it has caused to people's lives and their economies. Intersecting with this are other sudden changes including the eruption of demonstrations against the state in the United States, Hong Kong and elsewhere. Associated with these changes has been a lot of public reflection about the causes and what could have been done in anticipation.

One way some organisations have tried to cope with an unpredictable world is through the use of **scenario planning**. This is a body of practice that has been around since the 1960s. Scenario planning is not about developing accurate predictions of the future. Rather it is about identifying a range of major possibilities that might eventuate. Having identified such scenarios an organisation's challenge is then to identify how it might best respond to each of these possibilities, should they

happen. In their earliest forms, scenario planning exercises were very much expert led. But in the last decade or so participatory forms of scenario planning have become much more common, particularly in the field of natural resource management and the preservation of biodiversity.

There is a wide variety of approaches to scenario planning. But some characteristics of the process are more common than others. One very common sequence found in many scenario planning exercises involves: (a) the identification of various possible drivers of change which might be present or absent to varying degrees in the future, then (b) the examination of combinations of those drivers of change, some of which were more will be more plausible and likely than others. One common product of this type of exercise is a 2 x 2 matrix showing all the possible combinations of the two most important drivers of change. Each of these possibilities is then developed into an elaborated narrative description. This is what I call a "**drivers-first- narratives-second"** approach

The problem with this approach is that viewing the future in terms of just four possibilities sounds like quite an impoverished approach when we give more than a minutes thought to just how unpredictable events can be. Couched in more the theoretical terms, this approach fails on a criterion called ''***requisite variety***". This is the notion developed by Ashby in the 1950s, an expert in the study of control systems (otherwise known as cybernetics). In its simplest form, the idea is that a model of the world must have enough variety of possible states to mirror the states of the world we want to influence. This suggests a very basic evaluation criterion. When we are looking at the future, we should be looking at multiple alternative futures, not just one or two.

In early 2019 I contracted the development of a web application designed to enable participatory scenario planning online. This is called **ParEvo**. A core feature of the ParEvo scenario planning process is the development of multiple alternative storylines. These develop as a branching structure, starting from one seed but then developing in different directions, some of which again branch into alternative directions, and some of which do not. Another characteristic of this process is that these storylines are not developed all at once, but rather through a series of iterations. Existing storylines provide the constraints and opportunities for their extension, and those new extensions then provide further constraints and opportunities for further extensions. Developing narratives about the future, as seen during a ParEvo exercise, is **an incremental, adaptive, and exploratory process**. Not a once off event. (You can learn more about ParEvo works by viewing this YouTube video).

Built into the ParEvo process is the opportunity for participants to provide their own summary **evaluations of the storylines** that are generated. Participants are asked to identify which storyline they think is most probable, least probable, most desirable, and least desirable. These are the

current default evaluation criteria and are like those used in other scenario planning exercises. But the capacity exists for facilitators of a ParEvo exercise to edit these and use other criteria of their own choosing. When participants evaluation judgements are downloaded, they can be then displayed in a two-dimensional scatterplot. In the same downloaded dataset we can see three broad types of responses. Some storylines, usually only one or two, will have consistent desirability and/or probability ratings. Other storylines will have desirability and probability ratings which are contradictory i.e. some participants will have rated the storyline as most desirable, while others will have rated as least desirable. Or, some participants will have rated the storyline as most likely, while others will have rated as least likely. The third type of response are storylines which have not been identified as fitting any of the four criteria. They were not seen as most or least likely or desirable. Storylines in this group might be of two subtypes. Some of them might simply be relatively neutral in their characteristics, neither probable or improbable, nor very desirable or undesirable. Other storylines might simply be too difficult to assess on either of these criteria.

This brings me to what could be called **meta-criteria**. How do we assess the value of different criteria used to evaluate different scenarios? Such as desirability and probability, referred to above. The one criterion that I am giving the most attention to it the moment is usefulness, a topic of continuing and wide concern by evaluators. So, going back to the evaluation feature in ParEvo, how useful are these evaluation criteria? If they are not very useful maybe other criteria need to be found.

My current and very provisional thinking is that these two criteria are potentially useful. The reason why I think this is that there is some correspondence between these criteria and the distinction which economists have made between **risk and uncertainty**. There are various interpretations of this distinction – so, this is my provisional understanding of the difference. Risks are possibilities we are aware of and on which we can put a probability estimate. Uncertainties are possibilities that we are aware of but cannot make any probability estimate, and possibilities that were not even aware of, which by definition we can't make any probability estimates for. The distinction between risk and uncertainty is probably not black-and-white, our confidence in probability estimates probably exists on a continuum.

A connection can be seen here with the ParEvo evaluation criteria. Storylines with unambiguous most or least probable/desirable ratings fall within the category of risks. Storylines with contested most or least probable/desirable ratings could be seen as uncertainties. So perhaps to a lesser extent might be some of the storylines that did not get a most or least rating at all.

The main implication of this distinction between risk and uncertainty is that there will be some future possibilities that can be planned for and others which can't be. And somewhat paradoxically, we need to be prepared for both. From my reading so far, there are **two ways of responding to these possibilities**. For identifiable risks i.e. those with some clear and significant probabilities, specific strategies can be developed in readiness. Pandemic preparedness is an example, even though their recent implementation may leave a lot to be desired. For uncertainties surplus resources need to be available on tap, to enable new and additional responses to new and unexpected possibilities. James March, a famous organisation theorist, described these additional available resources as "organisational slack". This strategy is similar to an evolutionary strategy identified by ecologists, which is called "bet-hedging". Bet-hedging is a kind of response that is "*neither optimal nor a failure across all environments*" (Simons, 2011). A frequently cited example is the behaviour of annual plants which ensure that only a fraction of their seeds germinate in any given year. Slack is a potential source of innovation, but also of inefficiency. A decision not to use this strategy, in order to maximise efficient use of resources in the immediate present, can be a source of fragility when the unexpected happens. As was seen in the 2008 financial crisis, when many heavily leveraged businesses went bankrupt.

As already mentioned, the facilitator of the ParEvo exercise has the option of changing the default evaluation criteria. **Various other criteria have been proposed** as means of assessing the quality of scenarios. One of them is **traceability** - "*a process is traceable, means one can follow, what has been done and how a process came to its results*" (Kosow, 2015). Within a ParEvo exercise traceability exists in the sense that the specific contributor for each component of a storyline can be identified, unlike the texts generated by many other scenario planning processes (participatory and otherwise). But the thinking behind the content of a particular contribution is not immediately available – it could only be accessed by follow-up interviews with the contributor. At that point two aspects could be distinguished. One is the reasoning behind the choice of which existing storyline to add to with a new contribution. The other is the reasoning behind the content of that new contribution. These have not yet been investigated in a ParEvo exercises that have been completed to date but will be when the next opportunity arises.

A second and widely used criterion is **consistency**, defined by the absence of internal contradictions in the content of a scenario. Consistency has been described as a necessary but insufficient condition for a scenario to be plausible. Plausibility is a bottom-line requirement for contributions that are made within a ParEvo exercise. One writer has argued that "*consistency is understood as a safeguard against arbitrariness of scenarios. It is a substitute for empirical validation, which is not possible and not appropriate with respect to scenarios" (Kosow, 2015).* The downside risk of the

consistency requirement is the elimination of surprises from scenarios whereas in reality surprises are very much part of life.

A third criterion for evaluating scenarios is **polarity** (Missler-Behr,2002). Polarity means a scenario stands in significant contrast to another. The justification for valuing polar scenarios is that the combinatorial space that contains all possible scenarios is immense, so any sample of possible scenarios should contain as much diversity as possible. But the downside risk of using this criterion is that the more polar scenarios will appear like cartoon versions of reality i.e. over-simplified. In the ParEvo process three dimensions of diversity (derived from ecology) can be measured.

A fourth possible criterion for evaluating scenarios is **ownership**. In a ParEvo exercise it is possible to identify for each storyline who contributed to that storyline and to what extent. Storylines can vary in the extent to which they collectively developed by contributions from all participants, versus only developed by single participant, or a small clique of participants. In some circumstances it might be important that scenarios are developed which do have widespread ownership across the participants. But experience so far suggests that there are no grounds for assuming across all contexts that collectively constructed scenarios are going to be better in terms of probability, or desirability, for example. The relationship between ownership of storylines and other attributes of those storylines is not straightforward and needs further exploration.

A fifth potential criterion is **observability**. In the ParEvo exercises competed so far, storylines have varied in the extent to which the events could be reliably and easily identified, if they were actually happening. To some extent this varies according to how abstract the description is, but not wholly so. This criterion is similar to the data availability aspect of evaluability (Davies, 2013). It is an important one because it must influence how easily an organisation could recognise, and then respond in time, to events described as important possibilities in a ParEvo or other form of scenario planning exercise.

Any of the above alternative evaluation criteria can be built into a ParEvo exercise by its Facilitator. These options enable a significant degree of customisation, according to need. But they are unlikely to be exhaustive, even when just considered within the scope of the published literature on scenario planning. A second source of evaluation criteria is also available – the participants themselves. Their criteria can be identified using a particular type of **open-ended question** format, known as pile sorting or card sorting by ethnographers and website designers respectively. Participants in some of the completed ParEvo exercises have been asked to look at the completed storylines and to sort them into two piles, of any size, such that each pile has something in common which makes it different from the other pile. Participants responses are then documented, including both

qualitative description of what is unique about each pile, and a list of which storylines belong to each pile.

This option has only been used for a small number of ParEvo exercises so far. Some of the participants identified relatively straightforward **content differences** between storylines. Others identified what could be called **genre differences** between the storylines. In the most recent exercise (about the Coronavirus pandemic) these were some of the differences seen between the surviving and storylines:

- Stories which featured 'liberals as bogeymen' versus 'liberals as saints'
- Stories which were about 'everyone for themselves 'versus 'shared support '
- Stories were conflict was a prominent feature versus stories were conflict with less discussed or even absent
- Stories which prominently featured digital technology as a possible solution versus those that did not give it as much attention

These participant-identified differences have two **potential types of implications**. One is the way in which the participating organisation decides how to respond in future. For example, to scenarios involving conflict versus those which do not. The other is the way in which the participating organisation think about alternative futures i.e. metacognition. For example, the split between storylines featuring levels as bogeyman versus saints might suggest the need for less stereotyped thinking about the future.

There is further potential in the elaboration and refinement of the pile sorting method. For example, when participants are asked to sort storylines into two groups according to a significant difference, the kind of difference could be specified. Two particular types of difference seem relevant. One is significant differences in the *causes* of the events described, and the other is significant differences in the consequences of the *events* taking place. The first of these two could lead to the identification of what in other scenario planning exercises are already described as 'drivers of change '. But in the ParEvo context these would be identified inductively, after reflection on a variety of storylines. This would be a "narrative-first-drivers-second approach". This very flexible approach to the analysis of storylines stands in contrast to the narrowness and comparative rigidity of approaches which prioritise the identification of drivers of change as the first step. Other variations could also be explored, such as asking participants to pile sort the storylines according to significant differences in the *actors* involved.

So far, I have identified two types of sources of criteria for differentiating and evaluating futures**.** The first is largely expert and/or facilitator driven, though its implementation can involve participants making judgements on these criteria. The second is more open-ended and participant driven although enabled by an exercise facilitator. A third option exists, which would involve third parties as observers, those who were neither managing nor participating in an exercise. They could have an especially important role as identifiers of what was missing, at two levels. There may be whole types of storylines missing, and within the completed storylines there might be some actors, events or issues which were conspicuously absent. ParEvo now includes an option of allowing "observers", alongside active participants.

Making use of this third source of evaluation judgements might help the participants identify, what are for them, some of Donald Rumsfeld's famous 'unknown unknowns' and convert them into "known unknowns". These are the kinds of events that not only could we not assign probabilities to, but we did not even know about their existence. Once they are identified we can at least bring them into view to the point where we can think about how to respond to them, at least as a type of possible event.

Other information sources

ParEvo website: https://mscinnovations.wordpress.com/

https://mscinnovations.wordpress.com/introduction/ten-stages/#using

https://mscinnovations.wordpress.com/reconstructing-histories/

The Participatory Evolution Of Alternative Futures, Rick Davies, 2020 05 11 Paper submitted to a peer reviewed journal. 18 pages