Journal of MultiDisciplinary Evaluation
Volume 16, Issue 35, 2020

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Towards a Complexity Framework for Transformative Evaluation

Robert Picciotto
University of Auckland

Background: Complexity ideas originating in mathematics and the natural sciences have begun to inform evaluation practice. A new wave in evaluation history is about to break. A new mindset, new methods, and new evaluation processes are being summoned to explore and address the challenges of global pandemics, growing inequities, and existential environmental risks. This is part of a broader paradigm shift underway in science where interdisciplinarity has become the norm rather than the exception.

Purpose: This article explores the utility of a complexity framework for a more effective evaluation function. It unearths the antecedents of complexity thinking; explores its relevance to evolving knowledge paradigms; provides a bird's eye view of complexity concepts; uses the logic of complex adaptive systems to unpack the role of evaluation in society; and draws the implications of contemporary social challenges for evaluation policy directions.

Setting: Not applicable.

Intervention: Not applicable.

Research design: Not applicable.

Findings: The evaluation complexity challenge coincides with an urgent imperative: social transformation. The on-going pandemic has brought to light the disproportionate effects of health emergencies on disadvantaged groups and emphasized the urgency of improving the interface between humans and nature. It has also demonstrated the importance of modelling for policy making – as well as its limitations. Evaluation, a complex adaptive system, should be transformed to serve society.

Keywords: *complexity; computers; disciplines; emergence; modelling; paradigm, systems*

Introduction

Complexity ideas have invaded the evaluation literature. The number of journal articles devoted to complexity since the turn of the century has grown exponentially (Gerrits & Verweij, 2015). [1] The number of books addressing the same subject has increased as well. Thus, Forss et. al.'s edited volume (2011) addresses the evaluation of complex interventions and Wolf-Branigin's book (2013) focuses on the use of complexity concepts in social research while the overarching theme of Patton's *Developmental Evaluation* model (2010) is complexity and the subject index of his magisterial *Utilization Focused Evaluation* (2008) treatise includes 18 complexity entries.

This surging interest is part of a broader social research movement amplified by the advent of powerful computers and Big Data algorithms (Castellani, 2014). It also reflects the spreading notion that evaluation as currently practiced should refurbish its tool kit to break away from linear models and help improve evaluators' understanding of a volatile, conflicted and unstable word (Ramalingam & Jones, 2008). For example, according to Morrell (2021), complexity ideas are uniquely placed to inform evaluators' theory of change models focused on social transformation.

What explains this 'complexity turn"? Is it more than a fad? Are the complexity concepts originating in mathematics and the natural sciences transferable to evaluation? Will they generate new methods or only a new 'mindset'? Do they presage a paradigm shift in evaluation practice or merely repackage ideas that have long been explored by philosophers, social scientists, systems thinkers, and evaluation thinkers? How do they relate to the growing public disenchantment in societies characterised by global pandemics, grotesque inequities, and the existential risks of climate change and environmental degradation?

To generate debate about these issues, this article explores the utility of a complexity framework for a more effective evaluation function. It unearths the antecedents of complexity thinking; explores its relevance to evolving knowledge paradigms; provides a bird's eye view of complexity concepts; uses the logic of complex adaptive systems to unpack the role of evaluation in society; and draws the implications of contemporary social transformation challenges for evaluation policy directions.

The Antecedents of Complexity

Laplace's determinism, and Newtonian mechanics rest on rationalist, linear, and predictable laws according to which nature is made up of things embedded in other things—from elementary particles all the way to the planetary system. The seeds were sown by Copernicus in 1543 and by Galileo six decades later when he faced religious persecution after asserting, based on irrefutable astronomical evidence, that the sun did not revolve around the earth. [2] The fruits of his intellectual contributions were reaped when careful observation and experimental methods triumphed and the scientific revolution flourished.

The Clockworld Universe

By 1966, when the laws of gravitation were discovered, the notion of a 'clockwork universe' became compelling to Newton and his Royal Society colleagues. Even as they reiterated their belief in God, their linear scientific concepts helped to elucidate the dynamics of planetary motion. The machine metaphor of the universe subsequently light the fuse of the Industrial revolution in 1760. Work became specialized, standardized, and hierarchized, leading Smith (1776) to put forward the division of labour principle made famous by his vivid description of a pin factory.

Since then, rational empiricism has fuelled remarkable technological progress and, for

[1] The number of complexity related articles in the journal *Evaluation* trebled from 2001-2003 to 2013-2015.

[2] The Inquisition formally rejected his findings, banned his books, and condemned him to spend the rest of his life in house arrest, after two trials (1616 and 1633). It took more than three hundred years for the Catholic Church to clear him of wrongdoing.

centuries, challenges to the Cartesian view of the world remained dormant in the social sciences, in part because the advent of distinct disciplines each operating within its own silo segregated the physical, natural and social sciences from one another. But as trespassing across disciplines became more frequent, it became clear that the reductionist tenets did not jibe with chaotic dynamics, the evolution of biological complexity or even common sense observations about the natural world.

How do thousands of birds congregate in coherent flocks to avoid collision and find their way over huge distances without any overt central guidance? How do millions of ants create disciplined social groups able to design and construct huge and sophisticated nests to protect their tightly organized colonies? How did billions of independent economic agents, each focused on their own self-interest, generate rules that created modern markets?

The Roots of Complexity

The cultural roots of complexity theory are deep and diverse: the relationship between the whole and the parts is an integral part of indigenous cultures (Apgar, et. al., 2009). In the western intellectual tradition, the interface between the whole and the parts has long been explored by philosophers (Plato, Kant, Hegel). Equally, sensitivity to initial conditions concept lies at the core of the historian's discipline that Vico probed in his *New Science* (1725).

Complexity ideas also lurk in the early writings of Enlightenment philosophers. Smith's pioneering contribution to moral philosophy (1759) embodies philosophical tensions that persist to this day. While he viewed moral sentiments as an essential part of the human condition, he also conjured the invisible hand metaphor (1776), according to which freedom of production and consumption yield providential social outcomes—an early illustration of emergence and an emblematic example of the self-organising feature of complex systems.

Smith (1759) also presaged behavioural and social psychology when he described human beings as social creatures whose actions reflect the contest between their passions and such moral ideas as prudence and justice. Similarly, Hume (1711-1776) viewed selfish passions as the driving force of turbulent history while he associated peaceful progress with the sociability induced by commercial interests (Hirschman, 1977). The Marquis de Condorcet (1743-1794) observed the chaotic paradoxes of cyclical electoral systems where collective preferences are indeterminate.

Focused disciplinary views of the world are complementary. Complexity thinking was implicit at the creation of sociology. Comte (1798–1857) viewed the discipline he founded as the most complex of all sciences. His positivist stance is likely to have influenced Marx (1818-1883) who treated society as a system of interacting classes, Durkheim (1858-1917) whose functionalism foreshadowed phase transitions and Parsons (1902-1979) who formulated an action theory according to which society is made up of networks made up of interacting actors.

Arguably, an important moment came in 1890, at the very core of the clockwork universe, when Poincaré demonstrated that the problem of tracing the paths of three bodies in mutual gravitational interaction, while simple to pose, is impossible to solve precisely. In 1905, Einstein spectacularly demonstrated that matter is a domain of coherent energy storage. Finally in 1927, Heisenberg's uncertainty principle revealed that one cannot measure the position and the velocity of an elementary particle at the same time.

The Evolution of Knowledge Paradigms

In theory as well as in practice there is truth in Mencken's quip that "for every complex problem there is an answer that is clear, simple, and wrong."

More rigorously, Einstein stated that "it can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience"—a phrase aptly summarised by Roger Sessions thus: "everything should be made as simple as possible, but no simpler."

Undoubtedly, the straightforward experimental approach to science that emerged during the Enlightenment displaced superstition and obscurantist dogmas. The division of intellectual labor among the specialized academic disciplines that followed gave a powerful fillip to knowledge creation. The positivist approach to science, for all its limitations, fuelled rapid scientific progress and promoted extraordinary technological advances. Indeed, linear notions of orderly cause-effect relationships, and predictability of natural patterns, for example. still offer plausible explanations for a wide range of phenomena.

Nevertheless, the reductionist model of science is now widely perceived as incomplete. It only captures a shadow of the real world. Its basic postulate is erroneous: not all natural and human systems are complicated machines that can be disassembled in distinct parts to elucidate their functioning and ascend inexorably to knowledge of the complex. The mechanistic conception of the natural and social world is no longer tenable. It has been successfully challenged in such fields as biology, meteorology, epidemiology as well as the rapidly evolving fields of linguistics, cybernetics, communication, computer sciences, and artificial intelligence.

Life has its own rules that linear dynamics cannot always account for. As environmental constraints to unbridled economic growth became ever more visible, the wall that had long separated the natural and human sciences was breached and complexity took hold of the public imagination. Thus, there is ample support for Hawking's view that the twenty-first century is the "century of complexity". Knowledge creation must contend with a world that is inherently complex from the molecular to the global level. The world is not only complicated but also complex in the very sense evoked by the terms described in the next section.

The World is Complex

Embracing complexity has emerged as a plausible strategy for pushing back the frontiers of knowledge in our post modern age. Complexity thinkers have found their place in the scientific sun by transcending mechanistic methods. Their models have acquired momentum in the social sciences and it is time to put them to work in evaluation. A paradigm shift, i.e. a fundamental change in scientific concepts and practices (Kuhn,1962), is underway.

There is no turning back, e.g. epidemics, genetic defects, loss of biodiversity, and other natural and social processes share complexity characteristics. [3] These phenomena, and indeed the evaluation process itself, can be computer-simulated in instructive ways to transcend the dominant focus on the *how* question that social researchers have long been struggling with to also ask *why* through systemic simulations of complex processes, organizations, and contexts (Mikulecky, 2000).

Coming to terms with complexity is especially critical in contexts of rapid transformation. In the business world, Drucker was an early proponent of complexity. He stated that "every discipline has at its center today a concept of a whole that is not the results of its parts, and not identifiable, knowable, measurable, predictable, effective or meaningful" (Wood & Wood, 2000). He further opined that the Cartesian world view is static whereas the forces of innovation in well-run corporations have the edge in an 'age of discontinuity.'

While free of complexity terms, Latour's *Science in Action* theory (2005) has explored the actor-network interface in science, reflects the logic of complex systems, and throws new light on the structure of scientific revolutions. Specifically, he argues that seeking to explain how change takes place by simplifying reality through disaggregration of natural or human phenomena into distinct parts has paradoxically generated new, proliferating, intractable and deadly complexities: while human beings live in a closed system of restricted boundaries, they have behaved as if natural resources are infinite with potentially disastrous results because they conceived of the world as an open system without boundaries, as Galileo did (Latour, 1991).

[3] Financial crises are equally complex but the 2008 global financial meltdown also highlights the risks of excessive reliance on opaque, untested and simplistic computer algorithms.

Complexity Goes Against the Grain

Complexity concepts are not easy to grasp. The complexity nomenclature seems *prima facie* arbitrary, even absurd. Scepticism regarding its validity is widespread even though car driving or playing football demonstrate that the human brain can help us engage in extraordinarily complex maneuvers. Linear, mechanistic ideas of the world are comforting. We all yearn for simplicity. Our brains have limited processing capacities. Our short-term memory cannot accommodate large amounts of information. We need to filter the huge amounts of data we are bombarded with.

As a result, we systematically screen out unnecessary details and rely on routine, received wisdom and standard recipes to get on with our lives. We treasure predictability and the security it generates. We rely on experts and political leaders for answers. We avoid risky interactions. We dismiss unpalatable truths. We are creatures of habit even where changed conditions require us to learn and change. Human unwillingness or inability to adapt and come to terms with distant, low probability but catastrophic risks induces lack of preparedness and explains most accidents, business bankruptcies, etc.

Given the limited use of evaluation and the neglect of history, "an echo of the past in the future, and a reflex from the future on the past" according to Hugo, changes in scientific paradigms take time. The lengthy transition from one scientific theory to another requires consensus and repeated failures to falsify it while changes in how to carry out science is even slower given persistent attachment to methods and mental models that have proved their worth while the efficacy of new ideas and concepts have yet to be demonstrated.

The Interdisciplinary Imperative

The slow pace of paradigm change may also be interpreted in institutional economics terms: the specialized boundaries of the Cartesian science model generate information asymmetries that are costly to overcome. Thus, the *exit-voice-loyalty* trilogy of Hirschman (1970) offers an explanation for the durable faith in the clockwork view of the world: the resilient *loyalty* to a linear conception of science reflects an inclination to avoid or postpone *exit* from the modernist mental model in the hope that the exercise of *voice* within the confines of one's own discipline will help produce improved results.

This is consistent with complexity theory: exit is the positive *feedback agent* required for adaptation; voice is the *negative feedback agent* that restrains and postpones change; while *loyalty* regulates the resulting competition so that the resilience of belief is a complexity phenomenon: a manifestation of path dependence (Inglehart & Baker, 2000).

Economists view social systems as atomized and made up of self-interested individuals motivated by rational choice, self-preservation and economic advantage. Sociologists perceive human beings as highly receptive to the opinions of others, driven by collective protocols and sensitive to the signals emitted by leaders. Both perspectives have merit and they have been combined by economic sociologists and institutional economists who view human action as embedded in a web of information networks and social links, thus anticipating the advent of complexity thinking.

In a siloed, multi-disciplinary scientific world, incurring high transaction costs to break disciplinary barriers is an investment: the world is facing complex problems that no single discipline can solve on its own. Theories are rules of the game that can be played across disciplines. Instances of scientific crossover abound; for example, Maxwell's electromagnetic concepts helped construct digital computers while economics invaded the domains of psychology; sociology; law; geography; etc. with recent signs of imperialistic overstretch and robust backlash.

Evolution theory too has trespassed across disciplines: it has informed the work of epidemiologists, geologists and sociologists and it has been mobilized to explain changes within—and interactions among—the traditional academic disciplines (Cohen & Lloyd, 2014). In this way, disciplinary admixtures generate hybrid vigour akin to those bred into the new grain varieties that triggered the Green Revolution. Thus, interdisciplinary evaluation can help reform the policy environment and foster social change.

In the evolutionary economics field, the concept of *lock-in* has been fruitfully applied to show that switching costs, increasing returns to adoption, and network externalities generate a bandwagon effect so that small advantages buit into the initial conditions of a technology protect it from innovative alternatives. As a result, sub-optimal social outcomes materialize due to institutional inertia and resistance to change (Cecere et al., 2014).

Thus, and to close the circle, Darwin's evolutionary ideas have come to the rescue of complexity theory when Holland (1975) developed efficient genetic algorithms by mimicking biological evolution through mating and mutating programming solutions—a modern incarnation of the 'survival of the fittest' principle.[4]

The Evolution of Evaluation

From an evolutionary perspective, retracing the trajectory of the evaluation discipline is instructive prior to speculating about its future. A gradual, yet revolutionary, transformation has begun regarding what evaluation should examine; what kind of questions it should address; how these questions should be structured; how answers should be sought; and how the results should be interpreted. Indeed, such an evolution is imperative to help post-modern societies overcome the existential threats of the Anthropocene age.

All social actions are shaped by the shifting flows and ripples caused by changes in the intellectual climate. How then will evaluation be transformed? As an open, complex adaptive system, it will evolve in response to the signals of its operating environment; that is, to the ever-changing ideas and discourses that dominate decision making in the public sphere. Ideology is itself a complex phenomenon. Thus, political psychology scholars are studying belief systems by construing ideologies as conceptual networks of representations embedded in complex adaptive networks (Homer-Dixon et al., 2013).

Where the individualistic conception of society prevails, merit oriented and linear conceptions of evaluation dominate. This is the province of goal achievement evaluation, performance evaluation and randomised control trials. Where the relational view of society dominates, evaluation is worth-oriented and participatory and it emphasizes qualitative methods. Where mental models embrace admixtures of competition and cooperation in society, utilization focused evaluation and democratic evaluation models focused on significance and values come into play.

In complexity language, evaluation has evolved as a dynamic system toward a set of equilibrium states that prevail until a new ideological attractor emerges.

Grounded in the mental models and emotions of individuals, changing ideological frameworks act as *basins of attraction,* a set of states towards which evaluation spontaneously moves. Thus, evaluation reflects the unfolding balances that society strikes between competitive markets, governments and the civil society or local communities (Rajan, 2019).

Evaluation Waves

The inspired depiction of evaluation diffusion as a succession of waves embedded in larger tides of political ideology (Vedung, 2010) can be interpreted as *phase transitions* of a complex adaptive system: major changes in the structures and methods of evaluation took place in response to the ebbs and flows of alternating ideologies.

This initial wave of evaluation diffusion was *scientific.* It favored experimentation in line with the grand narrative of modern management and operational research. Its technocratic thrust sought to isolate public policy decisions from the messy world of politics. This instrumental yet idealistic model did not survive the anti-establishment cultural

[4] A genetic algorithm uses reproduction, crossover, and mutation concepts to help solve engineering problems by using a string of variables (such as size, shape, weight) to define and select design parameters so that they measure up to desired performance standards.

revolution of the late 1960's. As a result, a *dialogue-oriented* wave swelled in the 1970's.

Under this exceptionally innovative phase of evaluation history, stakeholders came centre stage in the evaluation process, promoted democratic decision-making and inspired progressive social action. Whereas positivist methods held sway during the scientific evaluation phase, the new participatory wave embraced constructivist and qualitative methods. However, and despite its reorientation away from the bracing experimenting society vision, evaluation remained a vocation rather than a commercial venture throughout the 1970's.

This conception of the evaluation occupation lasted until the 1980's when the participatory wave lost energy and was engulged by the *neo-liberal* flood of New Public Management (NPM) ideas. Through the exertions of business school graduates, evaluation was turned into a corporate management tool rather than an instrument of ethical governance.

Suddenly, citizens were relegated to the margins of society as mere consumers of government services and evaluators became hired guns at the service of vested interests. To back up a 'public choice' theory of government, an impoverished version of the evaluation discipline inspired by NPM ideas was mobilized, and evaluation mutated from a public good to a private good.

Around 1995, an *evidence based* wave rolled in. Evaluators are still surfing it. It remains sustained by the powerful pull of private interests in the public sphere. It is also characterized by an obsessive focus on the *'does it work?'* question (Stame, 2010). Randomization is back in vogue even though, following repeated skirmishes, the quantitative-qualitative paradigm conflict was uneasily settled through adoption of mixed methods (Mingers, 2004). In parallel, evaluation went global and 2015 was named the evaluation year by the United Nations.

By then, evaluation had become a commodity. By lodging commissioning in the evaluand sub-system; relying on external contractors to deliver evaluation services under tight control, NPM operatives cemented the influence of vested interests in the evaluation process. Will a new evaluation wave break as social pressures build up? Will a new evaluation agenda emerge under the influence of complex adaptation processes? Given the instability of the enabling environment this is hard to predict. Evaluation, a complex system, lies somewhere between order and chaos.

The Attributes of Complexity

Funnel and Rogers (2011) contrast the characteristics of complex systems with those of complicated ones: while both types of systems connote multiple interconnected parts, complex interventions cannot be properly evaluated through standard recipes; uncertainty clouds their potential outcomes; expertise does not guarantee resolution of the problems they raise; they are not replicable, etc. Other evaluation thinkers assert that complex systems lie somewhere between order and chaos; and that they are characterised by their situated status and their non-linearity (Forss et al., 2011).

From these perspectives, complex social interventions appear as adaptive, implemented in unpredictable ways by diverse agents, with results highly sensitive to context and initial conditions. [5] In the same vein, Patton (2008, 2010) lists properties associated with complex phenomena: high uncertainty; conflicting stakeholders' perspectives; large reactions to small actions; dynamic adaptation to changing conditions; self-organization among interacting agents.

A Bundle of Concepts

In all cases, evaluation thinkers describe complexity rather than define it. They are not the only ones to be perplexed: there is no universally accepted definition of complexity. A list of 31 names was compiled by Lloyd of the Massachusetts Institute of Technology (Hogan, 1995). The demarcation line between complicated and complex systems is fluid, and no consensus exists about how complexity can be measured.

[5] Complex social interventions address wicked problems; that is, problems that are difficult to solve given elusive, incomplete, contradictory, or changing definitions and requirements.

Alternative indicators include diversity of composition; the extent of 'surprise' in information content; the size of the computer program that describes the system; how difficult it is to construct; the ruggedness or cascades of detail of a system's fractal structure; the number of hierarchical subsystems (or building blocks) it encompasses, etc. (Mitchell, 2009).

Arguably, the meaning of complexity is fluid because of the fledgling nature of complexity research. According to Holland (2014), "we are still primarily at the stage of collecting and examining examples, much as was the case in the early stages of biology, or the early stage of physics before Newton, or the study of electric and magnetic phenomena before Maxwell … We are still a long way from an overarching theory of complexity". As things currently stand, rather than a single science or theory, complexity is no more than a bundle of concepts.

Models and Their Limitations

Not all complexity scholars agree that a unified theory of complexity should—or even could—be constructed. Scientists look at complexity as an emerging phenomenon to be analysed, while practitioners define complexity as an engineering problem to be tackled. Evaluation lies at the intersection of these two schools of thought: for philosopher Cilliers (1998), engaging with complexity entails addressing the specificity of individual systems that are contingent and cannot be adequately described by means of simple theories.

Emergent qualities frequently arise from complex interactions within systems and with their external environment. This allows for the possibility of understanding social phenomena by setting judicious boundaries around them, speculating about the determinants of their behavior, constructing non-linear mathematical models and interrogating them. While, beyond a prediction horizon, even the most sophisticated models do not deliver accurate projections, complexity thinking has laid the foundations for a coherent and promising scientific project.

Models necessarily simplify reality. They require assumptions, inferences, and input parameters: "all models are wrong, but some are useful" (Box, 1976). Model adjustments are frequently needed to secure a 'fit'. Thus, rather than establishing eternal truths, a model can only be empirically adequate, i.e. offer evidence that corroborates a hypothesis, falsifies a conjecture, or help guide further study. It does so by allowing simulation of diverse scenarios and examining the consequences (Oreskes et al., 1994).

Beyond the mathematics, the explanatory content of any model can be contested . This is to be expected: models are not reality but just as Mark Twain opined that 'history doesn't repeat itself, but it often rhymes", good models rhyme with reality and allow fruitful speculation about the 'what if' question. As Popper (1959) convincingly argued, all scientific statements are provisional, conjectural, and hypothetical. It is through iterative and contestability processes that the accumulation of knowledge proceeds as existing theories get falsified and new theories replaces them—until they are falsified in turn.

A Causality Revolution?

Causal reasoning research ranging from probabilism, manipulation, counter-factualism, and learning, for example, is alive and well but beyond philosophical concepts and theoretical breaktroughs, tool creation has always been an engine of scientific progress. Similarly, information science and the advent of powerful computers have opened a new chapter in evaluation history: theory-based models of social interventions can now be evaluated with far greater efficiency. This is in part because new mathematical notations have become available to eliminate the ambiguity between cause and effect of conventional algebraic equations. This is a genuine breakthrough: a very powerful instrument has been added to the evaluators' tool kit. Beyond association, it is now possible to address the causality question *(why?*) and the counterfactual issue (*what if?)* by turning complex theory of change diagrams into computable models, thus breaking free of the statistical and ethical constraints of randomization and field experiments and putting Big Data to work on social problems.

Specifically, using Big Data and the new algebraic notations, powerful computers can now help evaluators test the validity of program theories that direct the movement of the evaluand (conceived as a system) towards desired outcomes. Such theories may be embedded in the evaluand or replaced by alternative theories shaped by the evaluator with guidance from the traditional disciplines Thus, hierarchical Bayesian frameworks can mimic complex neural information systems to capture the qualitative dimensions of plausible intervention models (Pearl & Mackenzie. 2018).

Emergence

Some complicated mechanical systems (e.g., a steam engine) share the characteristics of complex systems in the sense that their whole is greater and different from the sum of their parts but they are not complex because the relationships that define them are largely linear. By contrast, complex systems are inherenly *non-linear*; that is, they evolve from initial conditions following rules that generate outcomes than are not driven solely by the parts which make up the system. This frequently leads to a lack of proportionality in the relationship between inputs and outcomes.

Emergence is evident when dunes are shaped by shifting winds and ripples of sand; when termites construct elaborate mounds equipped with chimneys and galleries; when hurricanes arise from interactions between wind, and warm surface waters; when cities arise spontaneously in seemingly adverse environments; when social networks deepen ideological fissures in society, etc. The resulting characteristic—*emergence*—is the hallmark of all complex systems: they display properties that their constituent parts do not have on their own.

Thus, evaluation, especially formative evaluation, displays emergence when its findings are utilized. As in other complex systems, emergence results from two-way *feedbacks* in which outputs are recycled to become inputs in ways that either reverse the change of some variable(s) in the system (negative feedback) or enhance it (positive feedback). Where positive feedback calls the shots, past events can weigh heavily on outcomes and the path is prologue (*path dependence*). This concept has been put to work in evaluations of technology markets, regional clustering, and organizations (Dobusch & Schubler, 2013).

Self-Organization

Emergent orders can arise spontaneously without external intervention (*self-organization*). The coherent behaviours of individual agents take place through indirect coordination. Their actions are often guided through the traces left by prior actions without planning, control, communication, simultaneous presence, or even mutual awareness (*stigmergy*). In human affairs, stigmergy is present wherever cooperation takes precedence over competition and it is therefore highly relevant to the evaluation of voluntary organisations.

Self-organizing, complex systems are omnipresent in human society as well as in the natural world, e.g. the human brain is akin to a complex system made up of 100 billion cells interconnected in complex ways to create perception, consciousness, and feelings. Similarly, the Internet is a self-organizing, non-linear, complex adaptive social system driven by 4.6 billion interacting agents who comply with simple rules to interact without central guidance and to exchange information through a huge global network made up of web pages—nodes, and hyperlinks—edges—(Rupert et al., 2006).

In such systems, feedbacks take place through *networks*, and between different levels of a hierarchy, i.e. a lower level in the system organisation influences a higher level which in turn may react causing new patterns to emerge. This reciprocal relationship is called *coevolution*, a term which also evokes the way organisms create their environment and are in turn moulded by it. Thus, evaluation of advocacy initiatives directs its focus on the co-evolutionary characteristics of policy regimes and advocacy campaigns.

System Parameters

The farther apart two system elements are, the less they influence each other. The final

equilibrium shape cannot easily be predicted. The outcome depends on two distinct system features:

1. *order parameters* that are internal and instruct parts of the system to cooperate and sustain the configuration or compete, leading to disorder; and
2. *control parameters*, that are external to the system and have the capacity to induce changes in the order parameters.

The relevance of these ideas to evaluation rests on its basic mission: to influence the control and order parameters of social interventions and policies to promote the public interest, as shown in the next section.

The degree of connectedness among the system components is embedded in a *correlation function* that determines how pairs of system elements influence one another and at what distance. The *correlation length* (set by the control parameters) is the distance threshold beyond which the system elements are free of the influence of other system elements. When the control parameter is tuned to the critical point, the overall connectivity of the system is hugely amplified.

Phase Transitions

The *tipping point* theory popularized by Gladwell (1963) describes how phase transitions often occur in society. In line with a power law, new ideas (i.e., signals) produced by innovative thinkers (*mavens*) are disseminated by a few well- placed individuals (*connectors*) whose networks of influence include charismatic individuals endowed with exceptional communications and negotiating skills (*salesmen*) in ways that generate a chain reaction.

A frequent feature of complex systems is *criticality* which materializes when a small incremental event triggers massive systemic change due to subtle interdependencies among the system constituents, e.g. a pile of sand will suddenly lose its conical shape and collapse after reaching a critical size once an extra grain of sand is deposited at its apex.

This occurs following an initial stage of accumulation which allows the sand pile to grow until the fateful additional grain causes a group of neighbouring grains to lose synchronization, a failure that spreads to other interconnected groups within the system and cause an avalanche. The threshold at which a phase transition starts is determined by the network configuration that connects control and order parameters.

This summons the threshold model of influence (Granovetter, 1983) according to which weak social ties in dense networks are exceptionally effective at inducing gradual behavioural changes since they are safer and free of the high transaction costs and risks to stability associated with the demandingness of strong ties.

A system reaches a critical *state* if the configuration of its components and the nature of their interrelationships make the system vulnerable following a relatively small input. The resulting change may be limited; for example, the collapse of the sand pile may be partial: some groupings of sand grains may fail to remain connected while others may resist cascade failure.

Thus, the outcome of a social intervention or a public policy may be highly significant and transformative—or it may be catastrophic. Phase transition ideas are at the heart of the feisty debates currently underway in the evaluation community with respect to sustainability assessments and social transformation. In *first order* phase transitions, the change is abrupt (e.g., liquid water into ice or steam) whereas in *second order* phase transition it changes continuously (e.g., magnetization).

Power laws relate the probability of a phase transition to the size of the transformation: there are far more small avalanches than huge avalanches, exponentially so. Many other systems behave according to the same *scaling l*aws: earthquakes, forest fires, power cuts, health problems, share prices, etc. In welfare economics, Pareto (1848-1923) discovered a power law according to which 80% of the effects come from 20% of the causes.

Boundaries

The boundaries of systems and their permeability characteristics that govern the influence of signals on their behavior lie at the core of complexity models. To allow analysis, *boundaries* demarcate the limits of a system's internal components and processes. Within them, a system has integrity, in the sense that its parts work together to generate outcomes that confirm the system's relative autonomy.

This is why a critical factor of evaluation quality has to do with how evaluations are framed. Systems are social constructions and the choice of evaluation boundaries affects not only the validity of evaluative conclusions but also how the evaluation will impact on society; that is, who may benefit or suffer from it (Williams & Iman, 2007).

Whereas *closed* systems are not affected by outside influences, *open* systems are. The boundaries of an open system are permeable so that open systems interact with their environment; that is, the other systems located outside its boundaries. This adds to the uncertainty generated by the sensitivity to initial conditions of complex open systems.

For some such systems (e.g., the weather), it matters a great deal:[6] extreme sensitivity to initial conditions make prediction of future states highly inaccurate if only because the inevitable error in measuring the initial state gets amplified as the system evolves. But for other systems (e.g., the solar system) non-linearity does not matter much,[7] and relatively accurate predictions are possible. This is also the case for some evaluands in the social world and in such cases complexity thinking adds little value.

Chaotic Systems

Complex and chaotic systems have *sensitivity to initial conditions* in common: owing to *nonlinearity*, [8] system states that operate under the same rules may nevertheless follow very different trajectories over time even if they are relatively close together at the start.[9] This property can be demonstrated through mathematical simulations of simple equations that yield dynamic trajectories that often seem totally random.

Some chaotic systems differ from complex systems: they involve fewer parameters and they are driven by simple rules that nevertheless produce highly intricate dynamics and seemingly random results. [10] Unlike complex systems, they are fully deterministic in the sense that their initial conditions can be defined precisely, and their behaviour is completely determined by pre-existing causes, but their end states are nevertheless highly uncertain.

The term chaos is informally and strictly speaking inaccurately used to refer to disorder and randomness, but some chaotic systems are predictable—for a while. They may eventually 'appear' to become disordered; that is, uncertainty of forecasts increases over time. Still, given their deterministic features, randomness remains confined. Yet, they tend to lack the self-organization and feedback features of complex systems that can allow a single steady or equilibrium state of behaviour to materialize.

---

[6] Chaos theory emerged because a rounded decimal number in a simple 12 variable weather computer model being tested by meteorologist Lorenz in 1961 yielded completely different weather patterns. This led him to famously ask "whether the flap of a butterfly's wings in Brazil can set off a tornado in Brazil".

[7] The solar system is chaotic but stable in human terms: it will take tens of millions of years (or more) for planets to begin shifting their orbits and a few billion years for planets to collide with one another.

[8] Nonlinearity means that causal links are more complicated than a single chain; for instance, they may involve feedback loops.

[9] In a simple vertical pinball machine, a ball is dropped to face two rows of equally spaced pins. The ball may end up in one of 16 possible pockets and it is hard to predict where the ball will land: the interaction with the first row of pins amplifies any minor difference between one try and the next, and the uncertainty is amplified by the interaction with the next row of pins.

[10] Chaotic systems generate *fractal* forms characterised by rough or fragmented geometric shapes that can be split into parts, each of which is at least approximately a reduced-size copy of the whole (exact or quasi self-similarity), They are pervasive in nature (e.g., clouds, snowflakes, mountains, river networks, cauliflower, broccoli, blood vessels).

Thus, complex systems operate in a region between order and randomness—where complexity is maximal. Dynamic complex systems may gravitate towards and settle into one or more possible steady states (*attractors*)—islands of stability in a chaotic world. Systems are considered robust when small changes in variables do not lead to highly disruptive changes because self-organization helps the complex system to adapt. [11] These are the dilemmas that transformative evaluation focused on sustainability must probe.

Evaluation as a Complex System

Evaluation is exceptionally well adapted to the facilitation of beneficial social change in a complexity age that requires interdisciplinarity (Scriven, 1991).

It is not only a discipline in its own right but also and uniquely, it is:

- a *transdiscipline:* it studies and helps improve certain tools of other disciplines, as does statistics or mathematics
- a *multidiscipline:* it deploys the methods and taps into the findings of whatever other discipline can throw light on the problem it is faced with
- a *multifunctional discipline*: evaluation is variously called upon to play many roles in its interactions with decision makers and stakeholders, as "arbitrator, scapegoat, trouble shooter, inventor, conscience, jury, judge or attorney"

The Scriven Trilogy Complexified

The most widely accepted definition of evaluation is deceptively straightforward: assessing the merit, worth and significance of things (Scriven, 2013):

- The *merit* dimension is intrinsic: it assesses the extent to which an evaluand complies with pre-determined goals, policies, norms, rules, and standards: *doing things right*.
- The *worth* criterion is extrinsic: it is about *doing the right things* for stakeholders.
- *Significance* is the bottom line: doing good as well as doing right from a public interest perspective.

Merit assessment examines the efficacy of *control parameters* located within the evaluand while worth assessment addresses the responsiveness of the evaluand to *external control* parameters. Together these influences help mould the *order parameters* that drive evaluand behaviour. Evaluation examines the feedback effects of a host of *relationships* within the evaluand boundaries (process) as well as those that connect the evaluand to its operating environment (context).

This is fully consistent with systems thinking and it complies with the tenets of scientific realism that seek to answer the overarching evaluation realist question: what works, for whom, in what respects, to what extent, in what contexts, and how? (Pawson, 2013). Of course, meeting all the dimensions of the merit and worth criteria relevant to an evaluand is rare. Tradeoffs are inevitable and there is no escape from collective action dilemmas.

Arrow's *impossibility theorem* (Maskin & Sen, 2014) states that if the preferences of two stakeholders or more need to be satisfied when choosing among three options or more then it is impossible to select goals that satisfy all stakeholders. How then can the evaluation process lead to evaluative conclusions? Significance, informed by ethics, is the end game when all pertinent data, findings and judgments regarding merit and worth are considered and when the size, importance, and transformative effects of the evaluand are synthesized to reach an overall judgment of value.

According to Fournier: "conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of

[11] Unlike merely complicated physical systems, robust complex systems can adjust to changes in the external or internal environment (e.g., removal of some parts of the system).

something). It is the value feature that distinguishes evaluation from other types of inquiry" (Mathison, 2005). It is precisely this feature that makes evaluation uniquely positioned to address the wicked problems that characterise our complex era. It is the acid test, the polar star that should guide all evaluators whether they report to the organization hierarchy in charge of the evaluand or not.

It is not enough to verify compliance with merit standards, and/or to assess the worth of an evaluand to the direct beneficiaries of an intervention. The public interest and duties of care towards the environment matter too and this where significance and value come in. These assessments cannot be made without reference to *ethics*—one more reason why evaluation is uniquely challenging: "Complexity is inherent in any ethical engagement, yet ethical frameworks are also models, and like all models, are limited, exclusionary, and incapable of accounting for the complexity of lived phenomena" (Woermann, 2012).

Power and Truth

According to Foucault (1977) "There is no power relation without the correlative constitution of a field of knowledge, nor any knowledge that does not presuppose and constitute at the same time power relations". In other words, there is no truth without power and power 'produces' truth through 'knowledge regimes.[12] Within such regimes, power uses co-optation strategies to defuse resistance.

If resistance is absent, power may encourage critical activity. So much so that resistance, even violent resistance, may be conjured, simulated, or even encouraged (while kept securely within bounds) if it does not arise spontaneously. Of course, once confrontation occurs, power will promptly seek to defuse it to showcase its domination and deter potential opponents.

Ultimately, power seeks normalization and conformity in ways that make forced compulsion and violence redundant through the logic of game theory: disproportionate power makes the exercise of power unnecessary (Hoy, 2004). Conversely, resistance is a vital component of knowledge regimes so that, paradoxically, the reality (or potentiality) of resistance validates the need for power.

However, this does not invalidate the concept of individual agency, dismiss the possibility of freedom or make principled opposition redundant. To be sure, power sets boundaries and restricts the space within which challenges to authority can operate. Still, evaluation is among the best placed among knowledge occupations to resist the excesses of power through collective action, feigned compliance with knowledge regime rules, subversive ideas grounded in evidence; and public exposure of the contradictions of power.

Evaluation is a Complex System

A key reason why evaluation has the potential to make a difference in the complexity age is that it is embedded in a system with permeable *boundaries*: unlike social research, it is not enclosed in an ivory tower. Indeed, evaluation is mandated to 'speak truth to power', even within organizations and/or in enabling environments that privilege vested interests and neglect the other dimensions of value.

Under the neo-liberal and evidence waves of evaluation diffusion described above, accountability has been measured mostly in terms of achieving organizational goals (*merit*) but achieving *worth* and *significance* is equally critical so that accountability for achieving socially and environmentally sustainable development outcomes should be the acid test of value for evaluands sponsored by organizations.

It follows that the extent to which principled, value driven norms are in place within the fabric of the *enabling environment* of society and/or in the *authorizing environment* of the organization in charge of the evaluand inevitably impact on evaluation outcomes even though the individual

[12] Fake news, alternative facts and the propaganda narratives propagated by the unregulated social media are extreme manifestations of this phenomenon (Picciotto, 2017).

evaluators tasked with evaluating—or helping to evaluate—an evaluand are mandated to observe these norms, whether they report to the evaluation user or not; that is, whether the evaluation is user-directed or evaluator-directed.

Accountability and Learning

Authority is an essential component of organization. It reduces the cost of information sharing within the organization and it coordinates the activities of internal agents. In its absence, chaos prevails unless order is achieved by consensus among internal agents (Arrow, 1974). In practice, both authority and consensus are present within organizations, in various degrees, and it is through the interplay of authority and internal consensus that change takes place in response to external shocks or gradually evolving operating contexts.

Some errors are unavoidable where uncertainty prevails but not otherwise. Minimizing unnecessary error is the crux of accountability and evaluation helps to make authority accountable. Hence, judicious adjustment of goals and internal protocols (control and order parameters) is an essential characteristic of effectiveness; that is, the organization should not only be accountable for the quality of evaluand goals and their achievement but also for their adjustment as operating circumstances change; that is, accountability extends to *accountability to learn.*

User-Directed versus Evaluator-Directed Evaluation

Evaluation has largely been conceived as a black box in the research literature. It is time to break it open and examine its contents before observing its actual workings. Once again, a complexity perspective helps: it conceives of evaluation as an adaptive system so that the effects of evaluation interventions are shaped not only by the quality of the evaluation—the main focus of contemporary evaluation research scrutiny—but also by:

- the *enabling environment* of society
- the *authorizing environment* of the organization that uses the evaluation
- the interactions between two distinct evaluation sub-systems:
  - the *user-directed sub-system* (A) and
  - the *evaluator-directed sub-system* (B)

Subsystem A is *always* present in organisations even if there is no formal evaluation unit within it (e.g., accounting, auditing, and other internal control instruments can be conceived as mild incarnations of evaluation). Equally, an evaluator sub-system B is *always* concealed within the authorizing environment (e.g., through newspapers; advocacy organizations, protest movements) even if no formal evaluation activities originate from outside the organization. In that abstract sense, A and B are always present and interacting. Hence, accountability and learning are two sides of the same coin and attesting to the effectiveness of control parameters embedded in an organisation is equivalent to ascertaining accountability not only for the relevance and significance of goals and results but also for organizational learning).

The extent to which the overall evaluation system has the potential to induce significant improvement in decision making in the public interest within a constantly evolving operating environment subject to periodic shocks hinges on the balance struck between negative and positive feedbacks (i.e., between stability and change) within the organization; that is, evaluation forms part of the bundle of *control parameters* that help mould the *order parameters* that prevail within the decision-making system.[13]

User-directed evaluation (sub-system A) has considerable value. It induces changes in order parameters to help achieve decision makers' goals (*single loop learning*); it may also

[13] The value of evaluation also depends on the quality of the evaluation; that is, the extent to which it accurately simulates the interaction between the evaluand and its operating context as well as the validity of its value assessment from the perspective of the citizenry and future generations.

probe the root causes of problems to help reconsider those goals (*double loop learning*). This strengthens the internal hierarchy and it promotes stability. On the other hand where a phase transition is required to face major changes in the operating context, restructuring of the internal mechanisms and rules that govern knowledge acquisition and system behavior may be necessary (*triple loop learning*) so that the hierarchy may be threatened and internal instability may result.

Whereas single loop learning and double loop learning strengthen hierarchy, triple loop learning challenges it and this may conflict with A order parameters that seek stability and may privilege authority. This means that user directed evaluation (A) occasionally needs the fillip provided by evaluator-directed evaluation (B); for example, when the organization is faced with the need for major internal changes in a phase transition, with instability implications that may be perceived as threatening in organizations that are not blessed with far-sighted leadership.

Thus, the main value of sub-system B lies in its independence from the possible vagaries of evaluation users, the potential capture of the organization by vested interests, and the noxious influence of such interests on evaluation norms and protocols—issues that may be too hard to tackle even by highly principled and persuasive evaluators embedded in sub-system A that is vulnerable to internal information biases and subject to internal resistance to change associated with the urge to protect organizational authority and promote internal stability.

On the other hand, sub-system B has limited value where all that is required is a gradual adjustment process, where the evaluation addresses the piecemeal social engineering interventions favored by Popper, or where management advice geared to a continuous re-alignment of internal agents' behavior towards achievement of relevant social goals is efficient and sustainable, as is the case in developmental evaluation contexts.

It follows that the 'weak ties' associated with user-directed evaluation are the most effective in relatively stable operating environments. On the other hand, evaluator-directed evaluation B may be critically needed to help ensure that internal order parameters are realigned in a timely fashion when the operating environment is highly unstable or affected by a phase transition that threatens the internal hierarchy.

Excessive influence of B over A through strict, continuous feedback undercuts the value of authority, shifts responsibility away from authority and ultimately reduces the autonomy of authority in ways that undercut accountabilty. Furthermore, the independent evaluation sub-system B suffers from information asymmetries with respect to the internal workings of the organization.

In sum, the best configuration in most circumstances lies in a judicious combination between A and B—where beyond its role in verifying the validity of individual self-evaluations and attesting to the validity of the order parameters embedded in A for the generation of evaluations in the public interest, B should be at the ready when called upon to facilitate a phase transition.

Table 1
User-Directed and Evaluator-Directed Evaluation Characteristics

	Risk of bias	Transaction costs	Potential for gradual adaptation	Receptivity to major phase transition
User-directed (A)	High	Low	High	Low
Evaluator-directed (B)	Low	High	Low	High

It follows that evaluation independence is a relative concept. Evaluators embedded in both sub-systems A and B are fully expected to demonstrate independence of mind and appearance, one of the many evaluator competencies required for high quality evaluation. But there may be limits to the influence over the organizational hierarchy that A evaluators enjoy so that the structural independence of evaluator-controlled evaluation B and their access to higher authority may have to come into play for optimum social outcomes.

Conversely, sub-system A evaluators will be prone to push back where their sub-system B colleagues put forward recommendations that fail to take adequate account of the internal user force field; that is, a Socratic dialogue ensues. Thus, both sub-systems A and B interact with each other horizontally as well as vertically within their organizational hierarchies (the authorizing environments of user-directed and evaluator-directed evaluations) as well as the overall enabling environment.

Evaluation Emergence

Evaluation is an intricate *network* where evaluative outputs are recycled to become inputs in ways that either reverse the change of some variable(s) in the sub-system (negative feedback) or enhance it (positive feedback). Typically, the higher level agents whose behaviors are moulded by the overall operating environment will direct the behaviour of the lower levels that in turn will react and cause new patterns to emerge so that the evaluation process is *coevolutionary*.

Consequently, and inevitably, the evaluation system is *emergent*: the evaluation whole is invariably different from the sum of its parts and evaluation outcomes are *non-linear* and hard if not impossible to predict. In turn the complex systems metaphor of the evaluation process implies a lack of proportionality between inputs and outcomes.

Evaluation costs are typically a minute fraction of those incurred by the organization; for example, 1-2% in international development institutions where evaluation is exceptionally prominent and well funded. It follows that the potential benefits that flow from a well conducted evaluation for a socially pertinent intervention are so high that it is plausible to assert that the overall evaluation enterprise is a high return venture even if it is rarely successful, just as the funding of start up companies, scientific research ventures, or the popular music industry, where a single blockbuster compensates for dozens of false starts to generate profitability.

Evaluation Model Configurations

Strictly speaking, evaluator-directed evaluation is not needed in high trust cultures where credibility does not rest on an arm's length relationship between decision makers and evaluators. Thus, in indigenous cultures, well-run voluntary organizations, and local communities endowed with abundant social capital, evaluation-directed evaluation is of limited value and even redundant given the strength of weak ties.

For example, both evaluators and decision makers have 'skin in the game' in Blue Marble evaluations which focus on interventions designed to address the major existential 'problems without passport' that face humanity. The same presumption applies in *developmental evaluation* contexts where evaluators and decision makers work together in alignment with the progressive values of an operating environment that reflects citizens' interests and promotes environmental sustainability (Patton, 2020).

What if the evaluation is commissioned by a progressive civil society organization in an enabling environment that opposes such values, a not infrequent situation where vested interests have a dominant influence on the enabling environment? Here again, one cannot expect independent evaluation lodged in the adverse enabling environment to be commissioned or, if it is, to have much impact. In such circumstances, user-directed evaluation can still be put to work through an *advocacy evaluation* approach. It evokes the transformational evaluation and culturally sensitive evaluation models currently on offer which are rarely used because as public goods they are underfunded in adverse enabling environments.

On the other hand, there are situations where the enabling environment is broadly democratic and propitious but where the

organization is captured by vested interests or has limited leverage and the evaluand is a highly dubious scheme (e.g., a large dam with limited power and irrigation benefits that silence rivers and displace hundreds of thousands farmers) that confirms the lack of appropriate self-evaluation standards In such circumstances, sub-system A is bound to fail and it is fully appropriate for ethical evaluators to adopt an *adversary evaluation model* (Picciotto, 2018) that implies resort to an evaluator-controlled evaluation mode (sub-system B).

Equally, when the enabling environment is undemocratic and the decision making organization is controlled by unethical vested interests (and therefore lacks a proper self-evaluation system A), only independent evaluation B is feasible—a *subversive evaluation* approach implemented in league with progressive local community organizations and civil society organizations. This conjures the radical stance of social constructionist, post-modern Fourth Generation evaluators.

Whenever the authorizing and/or enabling environments are adverse to the production of high quality, objective, no-holds barred evaluations, evaluators incur considerable risks and need far more support than they are receiving in the contemporary evaluation market society: there is a huge latent demand for adversary, advocacy and subversive evaluations.

In the real world, of course, the passions and interests of stakeholders come in bewilderingly complicated combinations; the enabling environment is a battlefield of ideologies vying for influence; and the distinctions drawn in the above narrative are often blurry and ambiguous. This is why an admixture of independent and self-evaluation is the most appropriate in a wide variety of contexts.[14]

Table 2
Evaluation Models and Operating Environment Characteristics

Model	Enabling Environment Propitious	Authorising Environment Propitious	User- driven evaluation (A)	Evaluator-directed evaluation (B)
Developmental & Blue Marble Evaluation	Yes	Yes	X	
Adversary Evaluation	Yes	No		X
Advocacy Evaluation	No	Yes	X	
Subversive Evaluation	No	No		X

Evaluation Use

The social value of evaluation rests on facilitating beneficial changes in the operating environment through evaluands created and managed by organizations. Evaluation influences the bundle of *control parameters* that help such organizations mould the *order parameters* that prevail within their boundaries. Assuming that an evaluation is of good quality, its value will depend on its impact on the evaluand from a public interest perspective which hinges on evaluation use—which in turn depends on the balance struck between the negative feedbacks of internal

[14] Advocacy evaluators evoke the myth of Antigone; the travails of adversary evaluators resemble those endured by Sisyphus; the quandaries faced by subversive evaluators are akin to the dilemmas that Heraclitus had to confront; as for developmental evaluators their work is never done in line with Penelope's experience; and finally evaluators torn between the demands of independent and self-evaluation are sailing against the wind, just as Ulysses did.

vested interests and the positive feedbacks of evaluation recommendations (i.e., between stability and change).

Why except under the developmental configuration do decision makers tend to push back and resist evaluation findings and recommendations and yet accept some of them? First, negative feedback is likely to dominate since pragmatic tolerance of some positive feedback can be indulged to protect path dependence and ensure overall system stability; that is, protect the hierarchical structures of both A and B.

Thus, a predominantly but not exclusively negative feedback will usually emerge to promote stability and allow gradual change.[15] On the other hand, in order to protect the legitimacy of hierarchy, it is often in the self-interest of decision makers to simulate fidelity to the principles of accountability and learning. They may opt to adopt some recommendations (positive feedback) while pretending that they are nothing new thus preserving the face saving asset valued by their internal stakeholders. A symptomatic illustration of this phenomenon is the paradox of 'obliteration by incorporation' (Merton, 1996). It makes a show of endorsing or incorporating a recommendation followed by amnesia about its source thus remaining free to implement the recommendation, now, later or never (obliteration).

Such intricate games can be modelled. Game theory has been used for the study of nuclear deterrence and war strategies and for rigorous aggregation of diverse stakeholders' preference functions embodied in mathematical utility functions (Shubik, 1998). Similarly, it could be used to explore the interactions between A and B and to discover better ways of evaluating the worth dimension of an evaluand by simulating the behavior of stakeholders and identifying the cooperative or non-cooperative games that yield stable and socially beneficial equilibrium solutions.

The Externality Gambit

The user-directed evaluation sub-system can take two forms: (i) internal (A1); or (ii) external (A2). Under both A1 and A2 options evaluation commissioning plays the role of a *control parameter*, endowed with the capacity to induce changes in the *first order parameters* that govern the evaluation process (terms of reference) and the *second order* parameters that govern the extent of evaluation use, i.e. the order parameters of sub-system A. Configuration A1-B is normally superior to A2-B since the latter incurs transaction costs associated with contracting and oversight, information asymmetries etc.

Why then is A2 it the most prevalent configuration in the market society? Because it may benefit from higher quality evaluation skills or better access to external information given internal staff limitations but also, in some cases and regrettably, as a clever stratagem that equates externality with independence so as to discourage the set up of an evaluator-directed evaluation function B or if this is not considered credible by an enabling environment that insists on complementary evaluator-directed evaluations for use as a protective buffer; an information filter or as a way to help defuse accountability for poor quality evaluations (plausible deniabilty).

The Imperative of Social Transformation

The evaluation complexity challenge coincides with an urgent imperative: social transformation. Given recurrent health emergencies, rapid environmental degradation, and the intense discontent caused by persistent racial discrimination, social immobility, huge inequalities, and the unmet promises of incremental policy changes in a world seemingly out of control, social transformation has become imperative.

The 2008 financial crisis was a spectacular demonstration of the huge risks to livelihoods associated with financial globalization. It came on top of secular trends characterised by

[15] This self-preservation strategy is frequently observed in the policing of prostitution, illegal gambling and the drug trade.

extraordinary increases in inequality, a perpetual war on terror, implosion of states in the Middle East and North Africa, and widespread popular rage and resentment (Mishra, 2017). In many western countries, epidemics of despair have been ravaging societies affected by stagnant wages, unequal access to health and education services and systemic cultural divides (Case & Deaton, 2020).

Nearly half of the world is still striving to subsist on $5.5 a day or less while the world's richest 1% have secured twice as much wealth as close to 90% of the world population. Whereas in the 1990's, democracy was on the march, the prevalence of liberal democratic regimes began to falter in 2005. By 2020, the aggregate *Democracy Index* compiled by the Economist Intelligence Unit (EIU) was the lowest recorded since the index began publication in 2006.

Fundamental changes in policy frameworks are therefore required to redirect current social trends away from runaway, inequitable and unsustainable growth towards enhanced human security and a fairer world. The political and social world and the natural world are closely intertwined. In turn this calls for evaluation methods and mindsets adapted to the dilemmas of a complex world facing potentially catastrophic systemic risks (Taleb, 2007).

The convergence of the Covid-19 pandemic and the global outrage caused by George Floyd's murder have come on top of the silent, insidious and deadly climate change crisis, adding up to a perfect storm. The pandemic has brought to light the disproportionate effects of health emergencies on disadvantaged groups and emphasized the urgency of improving the interface between humans and nature. It has also demonstrated the importance of complexity modelling and its limitations.

The new pathogen originated in response to increased human relocation to previously isolated areas, the unintended consequence of highly profitable factory farming that forced traditional producers to turn to the wild animal trade. [16] Racism, violence and the climate are not separate issues, e.g. indigenous peoples around the world are mobilizing against environmentally destructive projects funded by private interests. As a result, progressive forces are mobilizing against racial discrimination and the environmental movement has morphed into an environmental justice movement (McKibben,2020).

The Coronavirus crisis may be the tipping point that foreshadows another phase transition in the ideological environment that shapes public policy: this pandemic as well as others (e.g., Aids, Sars, Ebola) originated in animal populations affected by severe environmental pressures. It has become clear that their prevention hinges on reforming agricultural practices, discouraging factory farming practices and changing dietary habits; that is, on policies that acknowledge the relationship between human health and nature's health.

Towards Transformative Evaluation

Social transformation requires a qualitative change in the state of the world. Complex social systems are adaptive. They bring together agents that adapt as they react to one another within and among the levels of social hierarchies. This does not usually yield a stable equilibrium since agents keep changing their strategies in reaction to other agents' actions. But recurring patterns may be identified as the configuration of the overall system evolves.

The challenge of social transformation lies in achieving systemic change while avoiding social collapse. Uncertainty, the overarching characteristic of complex phenomena, is irreducible to linear conceptions of risk. It has a positive, possibilistic dimension that offers scope for action—and for progressive evaluation (Feinstein, 2020).

The evaluation community has begun to face the transformation challenge. It is refurbishing its assessment criteria. It is seeking to become more relevant, timely and technology savvy. It is tightening its

[16] Intensive livestock farming is also responsible for about 15% of carbon emissions as well as a fifth of transferable diseases

competency frameworks. It is shifting its focus from individual interventions to the higher plane of programs, policies and regulatory frameworks. A judicious combination of user-directed and evaluator-directed evaluation can help strike the right balance towards generating the major transformational changes required by the contemporary social predicament. But more needs to be done to transform evaluation.

Only then, will evaluation live up to the challenge of complexity and transformation:

- Evaluators will have to face reality: their occupation has been commodified and they need to break free from the tyranny of market forces. This will require reformed knowledge regimes, new evaluation governance systems, and diversification of funding sources.
- A formal commitment to a commonly agreed *professional ethos* should help distinguish evaluation from other occupations (auditors, management consultants and data scientists) that have captured evaluation concepts and distorted evaluation practice
- Evaluation should breach the wall between nature and culture. The relegation of rising inequalities and environmental stresses to side effects should be banned from the evaluation lexicon. Such existential challenges as global warming, pandemics and biodiversity extinction cannot be tackled unless the social and natural worlds are reconnected.
- The transdisciplinary and multidisciplinary mandate of evaluation implies a combination of general knowledge and deep specialization. Poor quality evaluation can destroy effective social programs or give credence to misguided policy interventions.
- Evaluation will have to make its way in the university world. Ensuring that evaluators are equipped with the knowledge, skills, and dispositions to exercise competent and independent assessments of transformative social interventions is a collective responsibility and controlled access to the practice through designation is a must.
- All contemporary models of professionalism stress the importance of self- management and autonomous control over occupational practices (Freidson 2001). Without professional autonomy, there is no way to tap economies of scale in administration, or avoid capture of the occupation by private interests or the state.

Conclusions

Complexity thinking embraces ideas that have long been explored by philosophers, social scientists, systems analysts… and evaluators. Putting complex adaptive system models to work will facilitate a much needed rapprochement between the physical, natural and social sciences. In turn, evaluators will have to adopt a new mindset and team up with other disciplines, including data scientists and mathematicians.

Hence, the complexity turn underway in evaluation is more than a fad. It is part of a fundamental paradigm shift in all the sciences. It roots are deep and the new science of causation associated combined with the availability of powerful computers will facilitate the integration of theories of change with interdisciplinary modelling grounded in the natural and human sciences. It will also break the monopoly of randomised control trials in evaluation by allowing computer based experiments that explore the 'why', 'what for', and 'what if' questions.

Finally, evaluators may find complexity concepts useful when they engage in meta-evaluation, an ethical imperative. Doing so may lead them to conclude that transformative evaluation implies a mix of user-driven and evaluator-driven evaluation, evaluation professionalization, commitment to a strengthened evaluation ethos, reformed professional governance structures, and new

funding mechanisms. Arguably, by embracing complexity ideas, evaluators may find it easier to live up to the ideals of their forebears.

Acknowledgements

This article benefited from diligent review by Eleanor Chelimsky, Osvaldo Feinstein, Linda Morra Imas, Jonny Morrell, and Michael Quinn Patton. They are not responsible for the errors and weaknesses of this version.

References

Apgar, J.M, Argumedo, A., & Allen, W. (2009). *Building Transdisciplinarity for Managing Complexity: Lessons from Indigenous Practice.* International Journal of Interdisciplinary Social Sciences, Voume 4, Number 5, pp.255-270

Arrow, K.J. (1974). *The limits of organization.* W.W. Norton, New York, NY and London, United Kingdom

Box, G. E. P. (1976), "Science and statistics" (PDF), *Journal of the American Statistical Association*, 71 (356): pp. 791–799

Case, A. and Deaton, A. (2020). *The Epidemic of Despair.* Foreign Affairs, March/April.

Castellani, B. (2014). *Brian Castellani on the Complexity Sciences*. Theory Culture and Society Website. October. https://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences/

Cecere, G., Corrocher, N., Gossart, C., & Ozman. M. (2014) *Lock-in and path dependence: an evolutionary approach to eco-innovations*, Journal of Evolutionary Economics, 24, pp. 1037-1065

Cilliers, P. (1998) *Complexity and Postmodernism* Routledge London United Kingdom

Cohen, E.B. and Lloyd S. J. (2014*). Disciplinary Evolution and the Rise of the Transdiscipline.* Informing Science: The International Journal of an Emerging Transdiscipline. University of Rhode Island, Kingston, Rhode Island. 17 189-215

Denzin, N. K. (2010). *Moments, Mixed Methods, and Paradigm Dialogs.* Qualitative Inquiry. 16 (6) pp. 419-427

Dobusch, L. and Schubler, E. (2013) *Theorizing Path Dependence: A Review of Positive Feedback Mechanisms in Technology Markets, Regional Clusters, and Organizations*. Industrial and Corporate Change, 22 (3), pp. 617-647

Feinstein, O, (2020). *Development and radical uncertainty.* Development in Practice. Routledge. Taylor and Francis. January 6 (online)

Forss, K., Marra, M., Schwartz, R. Eds. (2011), *Evaluating the Complex: Attribution, Contribution and Beyond,* Transaction Publishers, New Brunswick, NJ

Foucault, M. (1977). *Discipline and Punish: the Birth of the Prison*, Random House, New York, NY, p.27

Gerrits, L., and Veweij, S. (2015). *Taking stock of complexity in evaluation: a discussion of three recent publications.* Evaluation. Sage Publications, Thousand Oaks, California, 21 (4) 481-491

Gladwell, M. (1963). *The Tipping Point : How Little Things Can Make a Big Difference.* Back Bay Books. Boston, Massachussetts

Granovetter, M. (1983). *The Strength of Weak Ties: A Network Theory Revisited.* Sociological Theory. Vol 1, pp. 201-233

Hirschman, A.O. (1977) *The Passions and the Interests: Political Arguments for Capitalism before its Triumph* Princeton University Press, Princeton, New Jersey

Hirschman, A.O. (1970) *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States*. Harvard University Press, Cambridge, Massachusetts and London, United Kingdom

Hogan, J. (1995) *From Complexity to Perplexity*. Scientific American. June. Pp. 104-109

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*. U. of Michigan Press

Holland, J. H. (2014) *Complexity: A Very Short Introduction*, Oxford University Press, Oxford, UK, p.90

Homer-Dixon, T., Maynard, J. L., Midlenberger, M., Milkoreit, M., Mock, S.J.. Quilley, S.. Schroder, T., Thagard, P. (2013) *A Complex Systems Approach to the*

Study of Ideology: Cogenitive-Affective Structures and the Dynamics of Belief Systems, Journal of Social and Political Psychology, 1 (1) pp. 337-363

Hoy, D.C. (2004) *Critical Resistance: From Poststructuralism to Post-Critique.* The MIT Press, Cambridge, Massachusetts and London, United Kingdom

Inglehart, R. and Baker, W.E. (2000). *Modernization, Cultural Change and the Persistence of Traditional Values,* American Sociological Review. February, 65 pp. 19-51

Kuhn, T.S. (1962), *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press

Latour, B. (1991) *We have never been modern* Harvard University Press, Cambridge, Massachusetts

Latour, B., (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford. United Kingdom

Maskin. E. and Sen, A. (2014) *The Arrow Impossibility Theorem*, Columbia University Press, New York

Mathison, S. ed., (2005) Encyclopedia of Evaluation, Sage Publications, pp 139-140

McKibben, B. (2020). *Racism, Police Violence and the Climate are not Separate Issues.* The New Yorker. June 4. New York, NY

Merton, R.K., (1996). *On Social Structure and Science,* edited by Piort Sztompka. University of Chicago Press, Chicago, IL

Morrell, J, (2021). *A Complexity-based Metatheory of Action for Transformation to a Green Energy Future.* In Transformational Evaluation for the Global Crises of our Times, to be published by the International Development Evaluation Association (IDEAS), Exeter, UK

Mikulecky, D.C., (2001) *The emergence of complexity: science coming of age or science growing old?* Computers and Chemistry, Elsevier, 25. pp. 341-348

Mingers, J. (2004). Paradigm wars: Ceasefire announced who will set up the new administration. Journal of Information Technology. August. Palgrave. pp.165-171

Mishra, P. (2017). *The Age of Anger: A History of the Present*. New York: Farrar, Straus, and Giroux.

Mitchell,M. (2009). *Complexity: A Guided Tour.* Oxford University Press, Oxford, UK

Oreskes, N., Shrader-Frechette, K., Belitz, K., (1994) *Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences,* Science, New Series, Vol. 263, No. 5147 (February). Pp. 641-646

Pawson, R. (2013). *The science of evaluation: a realist manifesto*, London, SAGE Publications.

Patton, M.Q. (2008) *Utilization-focused evaluation (4th Ed).* Sage Publications. Thousand Oaks. CA

Patton, M.Q. (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. Guilford Press, New York

Pearl, J. & Mackenzie, D. (2018). *The book of why: The new science of cause and effect.* New York. Basic Books

Picciotto, R. (2017). *Evaluation: Discursive practice or communicative action?* Evaluation, July Issue 23 (3) pp. 312-322

Picciotto, R. (2019*) Is Adversary Evaluation Worth a Second Look?* American Journal of Evaluation, 40 (1) Sage Publications, pp. 92-103

Popper, K. (1959) *The Logic of Scientific Discovery, translation of* Logik der Forschung, Hutchinson, London, UK

Ramalingam, B. and Jones H. (2008) *Exploring the science of complexity: Ideas and implications for development and humanitarian efforts*. Second Edition. Working Paper 285. Overseas Development Institute, London, UK. https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/833.pdf

Rajan, R. (2019) The Third Pillar: *How Markets and the State Leave the Community Behind.* Penguin Press, New York

Rupert M., Rattrout, A., Hassas, S. (2008). *The web from a complex adaptive systems perspective,* Journal of Computer and System Sciences. Elsevier, 74, pp. 133-145

Scriven, M. (1991) *Evaluation Thesaurus. Fourth Edition*. Sage Publications. Newbury Park, London UK, New Delhi. p. 364

Scriven, M. (2013) *Key Evaluation Checklist (KEC*). July 24. p. 3 (footnote 8)

http://michaelscriven.info/images/KEC_7.25.2013.pdf

Shubik, M. (1998) *Game Theory, Complexity, and Simplicity Part 1: A Tutorial.* Working Paper 04-027. Santa Fe Institute. Santa Fe, New Mexico

https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/sfi-com/dev/uploads/filer/da/53/da53eb6a-a1e6-41dc-898f-f27bda9f7b97/98-04-027.pdf

Smith, A. (1759). *The Theory of Moral Sentiments*, D.D. Raphael and A.L. Macfie (eds.), Oxford: Oxford University Press, 1976.

Smith, A. (1776). *An Inquiry into the Nature and Causes of The Wealth of Nations, Representative Selections*, edited by Bruce Mazlish, Bobbs Merrill, Indianapolis, Indiana (1961)

Stame, M. (2010) *What Doesn't Work? Three Failures, Many Answers,* Evaluation, Sage Publications, 16(4) pp. 371-387

Taleb, N. N. (2007) *The Black Swan: The Impact of the Highly Improbable.* Random House, Penguin Books, London. UK

Vedung, E. (2010) *Four Waves of Evaluation Diffusion*, Evaluation, Sage Publications, 16: 263 pp. 263-277

Williams, B. & Imam I., Eds. (2007). *Systems concepts in evaluation: An expert anthology*. Edge Press of Inverness. Point Reyes CA

Woermann M. (2013) *The Ethics of Complexity and the Complexity of Ethics.* In: On the (Im)Possibility of Business Ethics. Issues in Business Ethics, Springer, Dordrecht, vol 37. p. 31

Wolf-Branigin, M. (2013). *Using Complexity Theory for Research and Program Evaluation.* Oxford University Press, Oxford, UK

Wood, J.C. and Wood, M.C. (2005*). Peter F. Drucker, Critical Evaluations in Business and Management. Vol.1.* Routledge, London UK and New York, NY

RP/rp

July 7th 2020