06/2013

University College London
Department of Security and
Crime Science

Dr. Noémie Bouhana

LEARNING FROM CRIME PREVENTION:

FOUNDATIONS OF A SYSTEMIST EVALUATION FRAMEWORK FOR ONLINE INFLUENCE ACTIVITIES

This page intentionally left blank.

Classification: Unclassified

Acknowledgements: This work has been supported by the UK Ministry of Defence. Any views expressed are those of the author and do not necessarily represent those of MOD or any other UK government department.

Table of contents

A note on intended use

This report is not intended as a practical guide to the evaluation of online influence activities. One of the main arguments presented here is that the design of effective and usable evaluation technologies requires end-user involvement. Rather, the purpose of this document is to inform, on the one hand, planners confronted with the issue of whether or not to undertake significant capability-building for influence activities, and, on the other, evaluation and operation designers faced with the task of improving the conduct of these activities.

Referencing

To streamline the reading experience, this report is not referenced as systematically as a document for academic publication would be. Nevertheless, significant portions of the following sections were synthesised from, or refer to, material authored by other scholars. Unless otherwise indicated in the text, this material is referenced as 'Further reading' at the end of each section.

Synopsis

Aims of the present report

The need for robust evaluation goes hand-in-hand with the undertaking of purposeful activity by hierarchical organisations with responsibility for multiple and competing tasks. Unfortunately, there is no such a thing as a standard design for the evaluation of human social activity, of which online influence activities (OIAs) are an example. Any evaluation design must be tailored to the specific activity under evaluation. To guide the appropriate design of evaluations, a framework is needed: a step-by-step process for operation and evaluation designers to follow. This report presents a rationale and blueprint for such a framework, by drawing upon the knowledge-base in crime prevention evaluation (CPE).

The challenges of evaluation

The problems facing the evaluation of OIAs are of four types:

- The problem of attribution: Evaluators must establish whether any change observed in the wake of an operation was caused by the operation's activity, rather than some other factor, as well as provide concrete measures of this change.

- The problem of generalisation: Evaluators have to establish how and why an operation succeeded or failed, in order to produce knowledge that will be of benefit to future operations implemented in different contexts.

- The problem of analysis: Evaluators have to keep track of (positive or negative) 'knock-on' effects, which accompany the implementation of operations in open social systems at different levels of analysis (e.g. individual level or community level).

- The problem of usability: Evaluation designs have to be usable given resources available to the organisation and the conditions under which operations take place. Robustness (the extent to which an evaluation design deals successfully with the first three problems) must be balanced against user needs.

In dealing with these challenges, it is worth looking to lessons learned in CPE. OIA evaluators can avoid retreading old ground, both in terms of questions already answered and mistakes already made.

Evidence-based vs. realist evaluation

Two competitive approaches to evaluation have emerged in CPE, as well as in other problem domains.

Evidence-Based Evaluation (EBE) privileges tackling the problem of attribution. It does so by advocating that evaluations should be designed on the model of Randomised Controlled Trials (RCT). A well-designed RCT will maximise the *internal validity* of the evaluation's findings, establishing with a high level of confidence whether the intervention was responsible, or not, for the changes observed.

EBE has been criticised for neglecting the problem of generalisation, inasmuch as RCTs mean to rule out confounding factors, such as individual differences and contextual variations. Yet these kinds of factors are likely to interact with the intervention in such a way that the same intervention implemented on a different population in another context will produce very different results. EBE is also criticised for neglecting the problem of usability. Carrying out an RCT is resource-intensive and requires the ability to impose artificial conditions upon the evaluated activity, which may be difficult to achieve when evaluating operations 'on the ground'. Approaches to influence activity evaluation which rely chiefly on gathering *Measures of Effectiveness* (MOEs) are closest in intent to the EBE approach, though they do not get close to its methodological orthodoxy.

By contrast, *Realist Evaluation* (RE) puts a premium on tackling the problem of generalisation. Realists argue that evaluation should not stop at establishing 'what works' by producing valid MOEs, but 'what works, for whom, in what circumstances', in order to strengthen the *external validity* of the evaluation. MOEs collected in one context will not provide grounds upon which to anticipate with confidence how an operation will fare if implemented in another. Realist evaluators strive to uncover the theoretical assumptions which underpin an intervention's design. They make educated guesses about the *mechanisms* which link the operations 'key ingredients' to the operation's outcomes. Without positing these mechanisms, the evaluator cannot draw lessons relevant to future operations. RE recognises that initiatives are implemented in open social systems, which are characterised by complexity and permeability. It is the realist evaluator's job to identify those systemic factors which interact with the intervention's activity to produce operational outcomes.

RE has been criticised for encouraging evaluators to eschew experimental methods, therefore not being able to establish that the intervention is responsible for change with a sufficient degree of confidence. Furthermore, while RE prioritises the generation of Context-Mechanism-Outcome configurations – which, according to the approach, make up the generalisable product of any evaluation – the notion of 'context' remains under-conceptualised, undermining the value of the approach as a robust and reliable evaluation framework.

A systemist approach to the problem of analysis

Systemism is an approach elaborated specifically in recognition that human social activity takes place within systems. These systems can be analysed with attention to their *composition*, *environment*, *structure*, and *mechanisms*, all of which are aspects of a system capable of

interacting with an operation's mechanism(s). Systemic analytical products, which articulate System-Mechanism-Outcome configurations could be used to strengthen a realist approach, substituting the ill-defined notion of 'context' for that of 'system'.

Given that operations can only be as good as the theories that drive them, the theories and models which make up the OIA knowledge-base should, like the operations themselves, be evaluated for fitness. Without a knowledge-base that is fit-for-purpose, IAs risk being irrelevant and ineffective at best, or counterproductive and damaging at worst.

Overcoming the problem of usability

Recognising that both of the mainstream approaches to evaluation – EBE and RE – have their limitations does not mean that any approach to OIA evaluation is condemned to choose between their respective limitations. These seemingly irreconcilable differences between EBE and RE stem from a category error: their proponents have historically treated evaluation as a scientific endeavour, when it is in fact a technological endeavour; hence, an engineering problem. While the goal of scientists is to maximise the internal and external validity of their research designs, the goal of engineers is to meet the requirements of the users of the systems that they design.

So far, evaluation has largely been treated as a scientific problem. The trade-off between the internal and external validity of evaluation designs implemented in complex open social systems has, therefore, appeared intractable. If, however, evaluation design is approached as an engineering problem, the task becomes one of eliciting requirements from users (OIA commissioners, designers and implementers) and setting out specifications for an evaluation design which balances optimally the trade-offs between these requirements.

Hence, there can be no one-size-fits-all or 'gold-standard' evaluation design. For any given operation, evaluators must produce a design which:

- delivers optimal measures of operational impact;
- captures operational inputs and outputs faithfully;
- eliminates the greatest number of impediments to the establishment of causal attribution (internal validity);
- provides an understanding of the processes involved in producing the outcome, including the circumstances in which these mechanisms are likely to work again (external validity);
- is usable in an operational environment.

In practice, such a design is likely to involve quantitative and qualitative mixed research methods, with element of both summative and process evaluation.

The need for an R&D programme

If evaluation is a technological endeavour, then effective OIA evaluation technologies – and, by logical extension, influence technologies – must be the product of a research and development (R&D) process. The technological rules produced by the R&D activity will address both the nature of problems and their solutions, and the tools and processes used to put these solution into action.

In the first instance, System-Mechanism-Outcome patterns uncovered by the evaluation will be formalised into heuristic rules. In the second instance, the rules will state which tools or means to employ to assist in the design and implementation of the solution. In the third instance, evaluation is likely, in the long run, to generate meta-technological rules about the effectiveness of certain classes of solutions against certain families of problems in certain families of systems.

A systemist evaluation framework for OIAs

A formative approach to evaluation is a flexible foundation for such an R&D process. Formative evaluations are conducted at the developmental stage of a new kind of activity, when there is insufficient knowledge about what sorts of operations might be effective, or even what, precisely, the activity is setting out to achieve. Formative evaluations encourage structured self-reflection about the nature of problems and ultimate objectives, as well as the development of innovative solutions. The process is very similar to requirements elicitation processes in systems engineering. It requires close collaboration between operation designers and implementers, and evaluators with an expert understanding of the influence knowledge-base.

Undertaking systemist formative evaluations is necessary to:

- elaborate well-posed problem statements (i.e. solvable problems, which are appropriately matched to tactical or strategic objectives);
- pinpoint and synthesise the relevant knowledge-base;
- elicit functional and non-functional requirements;
- set out specifications for influence technologies;
- generate explicit S-M-O configurations;
- delineate the pool of realistic interventions; and
- produce technological rules for the design of future evaluation and implementation tools.

The key steps involved in such a formative systemist evaluation process model are detailed in the penultimate section of the report, with plausible R&D activities suggested for each stage.

Conclusions

All four challenges of evaluation – attribution, generalisation, analysis, and usability – must be tackled if influence technologies are to become more efficient and more reliable. While it is possible to design functional, minimally-intrusive, MOE-based evaluations, this is not the approach to take when aiming for capability building on any significant scale.

Taking an engineering approach to the design of influence technologies will, however, lead to confronting a number of assumptions:

- It will challenge the idea that operation designers can go straight to theoretical models and empirical research in the basic or applied sciences – such as social psychology, social networking or decision theories – and put these findings to use without further ado.

  RECOMMENDATION: When commissioning, soliciting or turning to the products of research and theoretical development in the human and social sciences, users should keep in mind that these products need to be assessed against user requirements, both functional and non-functional, as would any other new technological system, prior to implementation.

- The notions of behavioural change which underpin current thinking on influence have their roots in thirty-year-old literature and need a significant update. Rigorous problem analysis and the subsequent synthesis of relevant knowledge-bases are likely to challenge the received wisdom that influence activities are and should be chiefly about changing 'attitudes'. A shift towards multi-level, integrated behavioural models is likely to take place, to reflect the state-of-the-art in the behavioural sciences. R&D activity may reveal ultimately that investment in basic science is required before influence models fit to drive OIAs can emerge.

  RECOMMENDATION: MOD should commission systemic syntheses of the literature on behavioural change in commensurate domains, which reflect the state-of-the-art in social environmental and ecological science, social cognitive neuroscience, and other systemic understandings of human behaviour, in order to generate new analytical frameworks for IO design, which do not rely outdated attitude-change models.

- Finally, a systemist outlook can only challenge expectations, if any remain, that IOs can achieve their objectives regardless of what goes on 'out there'. In open social systems, actions are as loud as words, if not louder.

  RECOMMENDATION: Building confidence in OIAs means, first and foremost, managing expectations of what they can achieve.

The next stage is to subject the blueprint of the systemist evaluation process model to further development, alpha-testing and fine-tuning, as a first step towards devising a coherent

R&D programme for influence technologies. Whether that course of action is desirable is not for the author to say. The present report can only aim to inform that decision.

Abbreviations

CBA Cost-Benefit Analysis

CP Crime Prevention

EBE Evidence-Based Evaluation

IA Influence Activity

IO Influence Operation

MoE Measure of Effectiveness

OIA Online Influence Activity

QE Quasi Experiment

RCM Rational Choice Model

RCT Randomised Control Trial

SCP Situational Crime Prevention

SECTION 1

Introduction: The challenges faced by evaluation activities

The ubiquitous expansion of the Internet and widespread diffusion of digital technologies has unleashed the potential for influence activities to be carried out in the cyber environment, as a complement to kinetic operations or on their own.

When purposeful activity is undertaken by a hierarchical organisation charged with responding to multiple and often competing tasks, the need for evaluation is unavoidable. Planners, operation designers, commanders and implementers seek answers to such questions as:

- *How do we go about identifying appropriate and achievable objectives?*
- *How do we go about designing an effective operation?*
- *How can the success or failure of the operation be convincingly demonstrated or measured?*
- *How do we explain success or failure?*
- *How sure are we that the operation was responsible for the changes observed?*
- *Would the same operation, implemented in another context, with a different target audience, produce the same outcome?*

> *"The emphasis of military operations is shifting more and more towards non-kinetic activities, such as Psychological Operations and Information Operations, which are geared towards influencing attitudes and behaviors of specific target audiences.*
>
> *Though many such activities are undertaken, there is little systematic evaluation of the effects they bring about and their effectiveness. As a result, it is not well known what these operations contribute to the overall operation and to what degree they are achieving their goals."*
>
> NATO, 2011

- *Would a different kind of operation (a cheaper one, perhaps) have done just as well?*
- *What transferrable lessons, if any, can be drawn from success or failure?*
- *Were there any unintended or unforeseen consequences?*
- *Could we have done more to anticipate these side-effects?*
- *In the long term, should this type of activity be a major or a minor component of the organisation's strategic portfolio of intervention technologies?*
- *Should it be abandoned entirely?*

Such questions have long been the concern of planners and intervention designers in domains as diverse as civil engineering, public health, commercial marketing, and crime prevention (CP).

In CP, the development of scientific evaluation framework has been accompanied by a corresponding improvement in the elaboration of theories, counter-measures, and strategies for crime reduction. As knowledge accrues as a result of evaluation activity, it also becomes possible to better motivate and justify the requisition of resources, with a greater degree of transparency and accountability.

Evaluation is the ultimate test of the ideas and techniques which underpin operations: *the more robust the evaluation process, the more robust the ideas and techniques which survive the evaluation hurdle, and – as evolutionary logic would have it – the more robust the operations.*

It is expected that a similar, virtuous feedback loop would accompany the improvement of evaluation practices in the domain of influence activities (IAs) generally, and online influence activities (OIAs) in particular.

Four key challenges can be identified at the outset, which must be overcome if one is to carry out efficient evaluations and design effective OIAs.

1. Evaluations must establish with confidence whether operations are responsible for the changes (outcomes) observed in the aftermath of IAs, as well as concretely measure these outcomes. This means that evaluations must produce solid evidence that success or failure is attributable to the operation being evaluated.

   This is the *problem of attribution.*

2. Evaluations must generate knowledge which contributes to the improvement of future operations, even if these operations are implemented in different contexts, target different audiences, or convey different messages than operations subject to evaluation in the past. This requires that evaluations not only produce solid evidence of failure or success, but that they uncover the reasons behind these results. If we know why a particular operation

succeeded or failed in a particular context, it becomes possible to adjust the design of future operations in light of that knowledge.

This is the *problem of generalisation*.

3. Operations are always implemented in 'open social systems', rather than in closed environments (such as a lab). Social systems are characterised by their permeability (they are acted upon by, and in turn act upon, other social systems). They are found at different levels of analysis (e.g. micro, meso, or macro level). To be useful to planners, evaluations must not only keep track of permeability effects (e.g. 'knock-on effects' of activity carried out in one system upon the components of another), but they must help us think through what will happen if an operation carried out at a micro level (e.g. influencing a handful of individuals) were implemented at another level (e.g. influencing a community).

   This is the *problem of analysis*.

4. Evaluation is not an abstract activity. Like the operations themselves, evaluations must be implemented: they are carried out by people with various levels of training, in often less-than-ideal circumstances, absent the kind of control over means of data collection and response elicitation that a scientist would otherwise take for granted. This means that, while evaluations must strive to be robust (i.e. they must strive to tackle effectively the first three challenges), evaluation frameworks and designs must also be sensitive to user needs (e.g. the reality 'on the ground'). Evaluation designs which are too difficult to put into practice will be discarded or improperly implemented by end-users and the evaluation activity will fail, regardless of how well it dealt with the issues of attribution, generalisation and analysis on paper.

   This is the *problem of usability*.

This report addresses each of these challenges in turn, through the lens of experience in CP. The intent is not to claim that the example of CP should be followed slavishly. Rather, the report highlights both the strengths and the weaknesses of CP evaluation frameworks, which are representative of issues faced by most areas of activity concerned with human change. The goal is to avoid rethreading old ground – both in terms of questions already answered and mistakes already made – in the course of developing an evaluation model for OIAs.

However, the report does not confine itself to a review of the 'state-of-play' in CP. Instead, it advocates combining the insights of competing approaches. Chiefly, it argues that while robust measures of effectiveness (MoEs) are necessary, they are not a sufficient component of evaluation. MoEs may tell us that something worked, not why it did. Yet establishing 'why' is a prerequisite to knowledge transfer between operations. As much as it needs valid MoEs, collected through robust designs, evaluation activity also requires sound theory, in order to provide grounds for *generalisation* as well as *attribution*.

Furthermore, solutions are proposed to address the shortcomings of CP evaluation frameworks, notably the neglected problem of *analysis*. The foundations of a formative evaluation process model for IO evaluation are laid out, rooted in an approach – systemism – whose main purpose is to structure our understanding of human action in open social systems.

Finally, it is suggested that the unproductive tension between those who would privilege attribution and those who believe that the chief purpose of evaluation is to produce generalisable knowledge can be overcome through a simple paradigm shift: While, traditionally, evaluation has been treated as a scientific endeavour, it is in fact an engineering problem. Evaluations, like influence operations, are technologies, not scientific projects. They should, therefore, be evaluated according to how well they satisfy user needs.

The entire problem-cycle of problem statement, problem analysis, operation design, implementation and evaluation must be subject to research and development, including, notably, the elicitation of user requirements. Only through this process can the last hurdle, *usability*, be overcome.

A systematic, fit-for-purpose, ambitious R&D programme must be devised if the capability to undertake strategic, effective and sustained influence activity is to be achieved.

Whether building that capability is desirable is beyond the scope of this report and left as an open question for the reader.

Structure of the report

Section 2 sets out the rationale for drawing from CP to inform progress in IA, by demonstrating notably that both domains face similar challenges. Section 3 provides an overview of the literature on IA evaluation and identifies outstanding, critical issues, which are addressed in the remainder of the report. Following a brief introduction to CP, Section 4 describes the evidence-based approach to evaluation, which aims to tackle the problem of attribution. Its relative neglect of the problem of generalisation is addressed by the realist evaluation framework, presented at lengths in Section 5.

While the realists' framing of evaluation activity in terms of the elicitation of context-specific, mechanism-based explanations is of arguable value to OIAs, their handling of the problem of analysis isn't robust enough to support operations in highly permeable systems, such as cyber environments. Section 6 introduces an analytical approach, systemism, which can address this shortcoming, while Section 7 sets out criteria to assess a knowledge-base that can best support the design of IOs. Section 8 describes the main families of evaluation designs in terms of their ability to tackle the four challenges of evaluation, making the case that none yet explicitly address the issue of usability, without which everything else is moot.

In Section 9, it is proposed that evaluation be recast properly as a technological endeavour, and that influence technologies should be the outcome of a research and development process, as are technologies in other domains of operation. The foundations of a systemist evaluation process model are laid out in Section 10. The report concludes in Section 11, where it is argued that strategic capability-building for influence activities will require the implementation of such an ambitious, systematic, and rational research and development programme for influence technologies.

SECTION 2

Why transfer knowledge from crime prevention?

A very short history of influence

Influence activities are socio-technical interventions which aim to change the behaviour of an individual, group or population, in support of a strategic, tactical or operational objective. IAs are, traditionally, carried out without recourse to the use of force or other means of coercion.

Historically, such activities have taken the form of loudspeaker or radio broadcasts and airborne leafleting by both Axis and Allied forces during World War II, as a means to sway the morale of enemy troupes or to spread factual information among civilian populations. More recently, influence operations (IOs) have been associated with the so-called 'Winning Hearts and Minds' strategy, rolled out alongside military action in the theatres of Iraq and Afghanistan.

A succession of recent asymmetric conflicts has driven home the need to win over allies and defeat opponents on the field of ideas, and to exercise soft power as often as military might. The established role of small groups of radicalised supporters in the resurgence of deadly terrorist campaigns on home soil and overseas has also highlighted the potential of targeted counter-influence operations for the purpose of intelligence-gathering, disinformation, disruption, and neutralisation of terrorist networks.

Harnessing the cyber environment

The cyber environment has been identified in successive government publications as a staging-ground for a new generation of threats to national security[1]. However, the internet is also a medium which can be exploited for the purpose of defence, to notable advantage.

In a world where strategic communication is often key to diplomatic success abroad and to securing popular support and material resources at home, the notion that versatile influence programmes could harness modern communication techniques and exploit the ever-widening reach of online social networking platforms has been gaining ground.

[1] See, notably, Cabinet Office (2011).

While traditional propaganda operations were often restricted to a geographical area, the internet enables messages to be conveyed regardless of international borders, straight into homes. Furthermore, the internet is a rich medium for information dissemination. It allows for immediate feedback, offers a range of methods for message diffusion (e.g. video, as well as text), and the opportunity for interaction and personalisation. Online delivery is also cost-effective compared to face-to-face interventions, with the added benefits of privacy and anonymity to participants[2].

Yet, although the internet is an attractive medium for influence, one of its chief benefits is also an important limitation: while messages can be disseminated remotely and reach widely, any impact of this diffusion may go unseen or be very difficult to attribute with any certitude.

This is one of the challenges of OIA evaluation: *how can we go about measuring and demonstrating the impact of influence activity carried out online?*

Old wine, new bottles?

Influence activity can be overt or covert, broad or limited in scope, and now, online or off. Whether the exercise of influence in a cyber-environment is an essentially different kind of endeavour from the exercise of influence offline isn't the main concern of this report, though the question must inevitably be raised.

The temptation can be great to jettison everything we know under the reasoning that anything 'cyber' must inevitably be novel, and that a new knowledge-base should be built from scratch. However, experience in other domains would suggest that people are people, and while environments change and technologies evolve, some rules continue to apply and recognisable patterns continue to emerge.

For example: while the transport revolution ushered in by the invention and diffusion of the automobile gave rise to new forms of crime, people continue to steal cars for broadly the same reasons that they used to steal horses, and they continue to prefer to commit their crimes a short walk from home.

While the particulars of problems and their solutions do change under the influence of the environment (e.g. RFID tagging has replaced horse branding), operating principles (e.g. property should be marked in some lasting way so that it can be tracked if it is stolen) remain.

In short: *one should try not to lose sight of the old wine for the new bottles.*

CP and OIAs face similar obstacles to progress

CP owes many of its methodological and practical advances to fruitful imports from the domains of clinical medicine and public health. Longitudinal studies, randomised controlled trials, and Haddon matrices have been successfully adapted to the investigation of the emergence of

[2] See Lustria et al. (2009) and Bewick et al. (2008) for a discussion of web-based interventions in the context of health.

criminal propensities in adolescents, the assessment of treatment effectiveness for serious offenders, or the development of prevention strategies in response to chronic episodes of sports violence, to name a few examples.

When a domain of activity faces enduring obstacles to progress, *it pays to turn to a neighbouring knowledge domain for inspiration or guidance.*

OIAs face a number of such obstacles, including, but not limited to: an open environment; a newly emerging topic; difficult access to quality data; limited technical expertise among practitioners; finite material and financial resources; and an underdeveloped scientific knowledge-base. As a field of scientific inquiry, the 'influence' domain remains unsystematic, largely conceptual, and fragmented.

This report is based on the premise that enough conceptual and technical areas of overlap exist between CP and IAs to justify drawing from the knowledge-base in CP, in order to inform the development of evaluation technologies in the influence domain. Areas of commonality, mainly in the guise of shared challenges, are summed up in Box 1.

Although CP continues to face obstacles to evaluation, academics and practitioners have been addressing these challenges and proposing solutions for going on four decades and can boast the benefit of some valuable experience.

Box 1 Commonalities between OIAs and CP

Practitioners in both domains grapple with many of the same challenges:

- They seek to influence individual, group and population behaviour for a specific purpose
- Their activity may be targeted at audiences who are (sometimes staunchly) antagonistic and opposed to the intermediate or the ultimate goals of the influence programme
- They operate in 'open systems', which are subject to the influence of other agents, groups and institutions
- They are involved in punctual operations, as well as large scale programmes which coordinate several smaller operations
- They need to demonstrate the effectiveness of their actions to secure further resources for action
- They are called on to measure intangible concepts in concrete ways (e.g. 'fear of crime', 'satisfaction', 'influence', 'attitude')
- They may be asked to carry out cost-benefit analyses, which translate success or failure in monetary terms
- They must minimize the unintended, negative consequences of their activity (e.g. 'problem displacement'), and try to maximise unintended, but positive consequences (e.g. 'diffusion of benefits')
- They want to learn from past activity to improve future interventions
- They want to apply these lessons in different contexts, against different kinds of problems, with different target populations
- They may have to work in partnership with staff from other organisations, institutions or agencies in order to implement the activity
- They have to convince collaborators and decision-makers that evaluation is a worthwhile undertaking
- They have to convey their findings to a non-specialist audience in an accessible way

In the next section, the state of the IA evaluation knowledge-base is briefly assessed, followed by an overview of those CP evaluation frameworks which have set out to tackle the problems of attribution and generalisation.

Further reading

Thomas, T.L. (2007). "Hezballah, Israel and cyber PSYOP." *IOSphere,* 31-35. Available from: http://fmso.leavenworth.army.mil/documents/new-psyop.pdf.

Murphy, D.M. (2012). "The future of influence in warfare." *Joint Force Quarterly*, 64: 47-51. Available from: http://www.ndu.edu/press/lib/pdf/jfq-64/JFQ-64_47-51_Murphy.pdf.

Keller, R.(2010). *Influence Operations and the internet: A 21st Century issue. Legal, doctrinal and policy challenges in the cyber world*. U.S. Air University: Air War College. Available from: http://www.au.af.mil/au/awc/awcgate/maxwell/mp52.pdf.

Pahlavi, P. C. (2007). "The 33 Day War: An Example of Psychological Warfare in the Information Age." *Canadian Army Journal*, 10:12-24. Available from: http://www.army.forces.gc.ca/caj/documents/vol_10/iss_2/CAJ_vol10.2_05_e.pdf.

Collings, D. & Rohozinski, R. (2006). *Shifting Fire: Information Effects in Counterinsurgency and Stability Operations.* Carlisle: US Army War College. Available from: http://www.au.af.mil/au/awc/awcgate/army-usawc/shifting_fire.pdf.

SECTION 3

Evaluating influence activities: A brief overview of the state-of-play

In search of a 'narrative of effectiveness' for IAs

A survey of the IA literature suggests that theory, practice and evaluation are still in their relative infancy – both in terms of the basic and applied science of influence, and of the development of protocols to implement and evaluate IAs. This relative under-development is even more acute when one considers OIAs specifically.

Arguably, the absence of a robust science of influence (an accepted corpus of well-validated theories of cross-contextual behavioural change) does much to contribute to a lack of confidence in the value of IAs.

As the case will be made later on, confidence in the effectiveness of a technology requires the availability of a narrative of effectiveness: a believable story of how and why the technology should work. Such a story is more believable if it 'fits' with already well-established stories. In other words, the narrative of IAs must fit with the most up-to-date scientific understanding of the processes that shape human behaviour – an understanding which, itself, should be supported by well-articulated theories and an accumulation of empirical evidence.

> *"During World War 1 the allies flew aircraft made of Balsa wood and fired archaic weapons across No Man's Land. In 2012 the allies fly super-sonic stealth aircraft and deliver precision weapons from unmanned drones.*
>
> *In World War 1 the allies dropped MISO/PsyOps leaflets. In Afghanistan in 2012 ISAF drops MISO/PsyOps leaflets. Unlike any other current military capability MISO/PsyOps has not evolved any substantial concept during the past 90 years."*
>
> McKay et al., 2012

Building such a narrative is a tall order. The science of human behaviour is an emerging, fragmented, and fast-changing field. Behavioural models can lack validation, often due to the difficulty in accessing large amounts of high quality social and human data.

In many fields (for example, criminology) several theories can compete, each seeming to provide part of the explanation, yet none standing alone. Concepts are often insufficiently defined. The problem is even more acute when one tries to synthesise knowledge across different disciplines, where words such as 'attitude', 'belief', 'motivation', 'disposition', 'influence', 'persuasion', 'perception' or 'intent' have different – at times incompatible – meanings.

Building a convincing narrative upon such uncertain foundations can seem a daunting task.

The limitations of current approaches to IA evaluation

In spite of the difficulty in building a narrative of effectiveness, a small number of more-or-less detailed frameworks or approaches to IA evaluation have been put forward.

These frameworks address many of same, or related, points which will be elaborated upon further in this report, though the terminology may vary.

This is unsurprising. Even a cursory overview of the evaluation literature across different fields concerned with human change – from public health, to education, to commercial and social marketing – will bring to light the same basic elements, which are intrinsic to the evaluation endeavour:

- an understanding of the different kinds of evaluations which can be conducted, and their respective purpose;
- an awareness of problems of causality and attribution (the demonstration that any given activity is responsible for the change being observed);
- the availability of data and the design of valid measures of impact;
- attention to the unintended consequences of the intervention; and,
- to varying degrees, articulation of the logic driving the activity.

Though most of these points are to some extent addressed, or at least acknowledged, by the literature on IA evaluation, a number of issues remain outstanding. It is argued here that these issues are critical, not only for the development of robust evaluation frameworks, but also for the successful implementation of IAs overall.

- **Conceptual fuzziness.** Many frameworks appear to take for granted that inducing attitude change is the main mechanism underpinning IAs, yet few, if any, offer an operational definition of 'attitude', nor discuss the need for clear constructs more generally, including in the case of concepts as essential as 'context of activity'.

Without clear operational definitions, it is impossible to establish with confidence which areas of basic and applied science should be drawn upon to inform operation design, or which MoEs may adequately capture evidence of impact.

- **A disorganised knowledge-base**. While some reports, notably RAND's, identify a social science knowledge-base which can serve as a guide for the design of influence operations (IOs), it does not establish *how* theories should be put to use to design IOs, nor upon which criteria that knowledge-base should be assessed.

  In other words, it does not make explicit the qualities that a knowledge-base must have to support the design, implementation and evaluation of IAs. Such criteria must be established clearly if the knowledge-base is to grow in an organised, rather than a haphazard, fashion.

  This lack of assessment criteria may explain why the current knowledge-base appears skewed towards rational choice models and attitude-change theories developed in the 1980s and 1990s[3], rather than reflect more recent developments in, for example, social cognitive neuroscience and human social ecology, which look at behaviour as the product of a situated process (person-environment interaction models)[4].

  Indeed, the most detailed of the frameworks (NATO 2011), acknowledges that the relationship between attitude and behaviour is scientifically contested and complex, but offers no guidance as to how address this – rather fundamental – conceptual weakness. Furthermore, while the NATO process model advises 'determining relevant contextual variables', it does not indicate on what basis this relevance should be established, other than, it seems, the analyst's 'common-sense'. Yet understanding and monitoring the impact of 'context' is likely to be crucial to any IO, especially those carried out in the cyber environment.

  An understanding of the place of basic science and applied scientific knowledge in the cycle of activity development, as well as explicit selection criteria, would progress the influence knowledge-base beyond its outdated neglect of contextual factors and interaction effects, as well as provide guidance for problem analysis and solution design.

[3] More recent literature is very much circumspect and nuanced about the causal relationship between attitude and behaviour. See, notably, Glasman and Albarracín (2006) for a meta-analysis of the factors which impact the attitude-influence relation.

[4] This state of affairs finds echo in the area of offender profiling, where reliance on outdated trait psychology has undermined confidence in the utility and reliability of behavioural and crime scene profiling as an investigative technique. See, for example, Alison et al. (2002).

- **Lack of integration with operation design**. The extent to which the evaluation activity should be integrated with the operation design activity, and the mechanisms through which the former should feed back into the latter, are not clearly articulated. Yet it is through these mechanisms that evaluation technologies can be counted on to improve operational effectiveness, taking IOs beyond mere 'craft'.

  Taking the NATO framework again as an example, operations designers are advised to identify sources of data and evaluations methodologies *before* selecting a form of intervention, which seems to run counter to the logic of evaluation design in most other domains (i.e. the design, including data and methods, is selected to fit the operation, not the other way around).

  More crucially, the quasi-exclusive focus on developing MoEs fails to address the issue of knowledge transfer (the problem of generalisation). The MoE approach fails to recognise that evaluation activity must capture operation-context interaction effects. It is not clear what can be learned from the 'MoE score' of an IO carried out in one context, which can be turned into a 'lesson learned' for an IO carried out in another.

- **Neglect of implementation**. Although approaches to IA evaluation emphasise the need to improve the development of outcome measures, the same attention is not devoted to the monitoring of implementation and the development of output measures. Integration of these implementation measures into the product of evaluation activity is neglected. Yet, experience in CP shows that evaluating operational implementation as well as overall operational effectiveness is a core component of evaluation activity.

- **No long-term, capability-building programme**. The open source literature reviewed for this report outlines general principles for the conduct of evaluation, but it doesn't offer guidelines for a long-term development programme. Yet, to build substantial confidence in IAs generally, and OIAs in particular, research and development processes and targets need to be set out.

These issues are addressed in the remainder of this report. The next section provides a very brief introduction to CP evaluation and the approach devised to deal with the problem of attribution.

Further reading

NATO (2011). *How to Improve your Aim: Measuring the Effectiveness of Activities that Influence Attitudes and Behaviors.* RTO Technical Report. TR-HFM-160. Available from: http://info.publicintelligence.net/NATO-MeasuringInfluence.pdf.

Larson, E. V. Darilek, R.E., Gibran, D., Nichiporuk, B., Richardson, A., Schwartz, L. H., and Quantic-Thurston, C. (2009). *Foundations of Effective Influence Operations: A Framework for Enhancing Army Capabilities*. RAND Corporation. Available from: http://www.rand.org/pubs/monographs/MG654.html.

Perry, R.L. (2008). "A multidimensional model for PSYOP measures of effectiveness." *IOSphere*, 9-13. Available from: http://www.au.af.mil/info-ops/iosphere/08spring/iosphere_spring08_perry.pdf.

Hutchinson, W. (2010). "Influence operations: Action and attitude." Proceedings of the 11th Australian Information Warfare and Security Conference, Edith Cowan University, Perth Western Australia, 30th November - 2nd December 2010. Available from: http://ro.ecu.edu.au/isw/33/.

Rowland, L. and S. Tatham (2008). *Strategic Communication & Influence Operations: Do We Really Get It?* Special Series. UK Defence Academy. Available from: http://kingsofwar.org.uk/wp-content/uploads/2010/09/RowlandsTathamInfluencepaper1.pdf.

Rate, C.R. (2011). *Can't count it, can't change it: Assessing influence operations effectiveness*. Carlisle, PA: U.S. Army War College. Available from: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA560244.

Munoz, A. (2012). *U.S. Military Information Operations in Afghanistan: Effectiveness of Psychological Operations 2001–2010*. Santa Monica, CA: RAND, National Research Defense Institute. Available from: http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1060.pdf

SECTION 4

Evidence-based evaluation: Tackling the problem of attribution

The crisis of confidence in CP

The notion that CP efforts should be evaluated scientifically gained traction as a result of the crisis of confidence which CP initiatives, notably community policing activities undertaken by law enforcement units, suffered in the 1970s and 1980s.

Bolstered by data from official crime statistics and victim surveys, which seemed to show that few, if any, crime reduction initiatives had any effect at all, the argument that 'nothing worked' in CP – be it policing, incarceration, offender therapies or ambitious social change programmes – started to gain strength. Not only was confidence in law enforcement and the CP agenda waning, but so was the belief that criminological research could ever produce knowledge that would lead to concrete reductions in crime and criminality.

This pessimistic view was eventually challenged in the 1980s and the early 1990s, a period which heralded three decades of intense development in policing and CP initiatives, with the introduction of Problem-Oriented Policing (POP), 'hot-spot' policing, Crime Prevention Through Environmental Design (CEPTED), Design Against Crime, and, last but not least, Situational Crime Prevention (SCP).

The new approaches shared two broad tenets: first, that CP initiatives should tackle discrete and well-analysed crime problems, rather than broad social issues; and, second, that they should move beyond targeting *criminality* towards targeting *crime.* In other words, less effort should be spent trying to prevent the emergence of criminality among the population (through social programmes and general deterrence) or reform career criminals (through various approaches to treatment), and more on disrupting the immediate causes of crime events (for example, by removing opportunities for crime, such as through target hardening).

Proponents of the 'new wave' reasoned that disrupting the immediate causes of crime events should have near-immediate effects. Not only would crime be reduced quickly and concretely, but reduction would be more easily measurable, compared with programmes whose diffuse effects might be years in the making.

From 'gut-feeling' to science-backed practice

While this shift in CP thinking was taking place, a new era of management was taking hold in the governance sphere, characterised by a target-oriented ethos. The performance of the police was to be assessed through the ups and downs of crime statistics, in a way it had not been before. In the wake of these changes, a new culture spread, which privileged knowledge-based activities over experienced-based practice, and considered that success (or indeed failure) should be objectively measurable.

Not all of the CP approaches introduced during that period met with the same success on the ground. Nevertheless, the idea that CP should, from then on, be based less on gut-feeling and anecdotal evidence, and more on science-backed practices took hold.[5]

Evaluation became an exercise in the demonstration of effectiveness, whose ultimate aim was to inform programme management practices, as well as rationalise investment in public programmes, and, importantly, improve the overall CP knowledge-base.

The rise of evidence-based evaluation (EBE)

In this endeavour, two (avowedly competitive) schools emerged, which sought to overcome the unproductive pessimism of 'nothing works,' in favour of the more pragmatic question of establishing 'what works'.

The first systematic evaluation tradition to be established in CP emerged out of the 'evidence-based policing' movement championed by Lawrence Sherman and others. This approach draws its philosophical and methodological inspiration from the practice of evaluation in medicine.

As is best-practice in the medical field, evidence-based evaluation (EBE) puts the highest premium on establishing *internal validity*. In other words, it seeks chiefly to establish *with the highest level of confidence achievable* whether an initiative is responsible for changes observed, as opposed to something else. Above all else, *EBE sets out to tackle the problem of attribution.*

Given inherent difficulties in establishing attribution, EBE advocates the use of the strongest research designs, the most rigorous analytical methods, and the collection of the best quantitative data available.

Opting for a tightly-controlled experimental approach allows the evaluator to 'confound' (rule out) other possible sources of influence. The randomised-control trial (RCT), which involves the careful selection of study participants and their random allocation to treatment and control groups, is considered the gold-standard in evidence-based evaluation design. While exacting to implement, RCTs provide the best measure of attribution and can be expected to produce the best scientific evidence of 'what works' (and what doesn't).

IA evaluations which focus on the development and application of measures of effectiveness (MoEs) come close, in spirit, to EBE, though EBE proponents would criticise the

[5] The bulk of the preceding discussion is drawn from Ellefsen (2011); see also Braga & Wiesburd (2006).

lack of experimental methodology. Nevertheless, in both cases, *the main objective is to capture a quantitative a picture of 'what worked'.*

The 'realist' challenge

What else could one demand of evaluations than they use the most advanced research methods available and produce the best possible scientific evidence, putting solid numbers to often nebulous effects?

Yet another group of scholars began to question whether EBE's almost slavish focus on internal validity didn't come at too high a price: the sacrifice of *external validity*.

If internal validity measures the extent to which the treatment is responsible for the outcome, external validity captures the extent to which findings are generalizable (i.e. the extent to which findings are relevant in settings other than the given evaluation setting).

In other words, critics argued, it is not enough to say that the intervention was successful in the case under evaluation; one wants to be able to say with some measure of confidence *whether the same intervention would be successful elsewhere.* One of the main purposes of evaluation, after all, is to inform future action, not just to assess current efforts.

As Lieberson and Horwich (2008:18) highlight:

"Granted, a well-executed randomized experiment provides the social researcher with a strong basis for causal inference [attribution of the cause(s) of an outcome to one thing or another]; but even then, a second issue is the broad range of possible conditions that operate to affect the specific results from such an experiment. For example, in the case of the effect of a training program, the experiment can tell the researcher what the effect is of the specific training program on the specific subjects in a specific location. A wide variety of experiments would be needed to work out [the broader range of] conditions [which could affect the results]."

In other words, establishing external validity involves replication: conducting the same RCT at different times in different circumstances. The problem arises, of course, as to what to do with the results when, as is often the case in CP (and elsewhere), an intervention that seems to work in one situation encounters less success, or altogether fails, in another.

More than just 'what works'

Critics of EBE, chief among them Nick Tilley, advocated a shift to a new perspective. CP initiatives, they pointed out, are not carried out in the controlled environment of a laboratory, but implemented in messy, natural conditions.

Day-to-day, CP activity doesn't follow the strict guidelines of a scientific trial. And even if it did, it is doubtful that the experimental assumptions inherent in RCTs could be met in the kind of natural setting (e.g. prison, court, gang, neighbourhood) where crime reduction activity usually takes place. Despite its proud scientific heritage, the 'gold standard' of RCT might not be

suitable when evaluating interventions of a social nature[6]. Imposing robust but tightly controlled methodologies on evaluations might produce reliable results, but it cannot tell us much about how the initiative would 'behave' if it were rolled out and 'routinised' in a variety of environments and upon a diverse population.

Hence, the realists claimed, evaluation should not stop at establishing 'what works'. To be of use to planners, it should establish 'what works, *for whom, and in what circumstance*'.

Tackling the problem of attribution is all well and good, but what about *the problem of generalisation*? Thus, the school of realist evaluation was born.

Positivists vs. realists: A clash of scientific philosophies with implications for the 'real world'

The (often lively!) debate between proponents of EBE and defenders of realist evaluation reflects an old schism in the philosophy of science, which opposes positivists and realists. Rhapsodising about matters of philosophy is far beyond the scope of this report, and likely of limited interest to its intended audience, but it is worth mentioning this disagreement, which is at the root of clashing visions in many domains of human action.

Facing-off are the *positivists*, who deem that scientific knowledge amounts to observables (data), and *realists*, who hold that scientific knowledge is made up not only of observations (of which experimental results and MoEs are examples), but also of unobservables, such as theories, hypotheses and causal mechanisms.

Realists require not only *observations*, but also *explanations*.

This schism is more than mere academic dispute and manifests concretely in ways that impact social interventions.

For example, in CP the so-called 'risk factor' approach is positivist by nature. It involves the statistical analysis of population characteristics in order to identify factors associated with criminality. These factors are, in turn, used as a measure of an individual's 'risk' of future involvement in crime. Checklists of 'risk factors' (e.g. for delinquency; for radicalisation) are popular among decision-makers: design a scorecard and one can quantitatively assign individuals to one category of interest or another.

But this is not without downsides. Since statistics are about correlation, not causation, the factors uncovered through this approach could be any number of things: a predictor, a symptom, a statistical accident, or, if one is very lucky, a cause. A 'risk factor'-based framework doesn't explain how or why a given factor (e.g. personality trait; education level; attribute such as gender or age) is associated with criminality (or any other outcome); it can only say that it is.

[6] Note that this critique of RCT isn't exclusive to CP. It is also present in the domain of community-based preventive medicine, which involves large-scale programmes aimed at whole populations. See, for example, Rootman et al (2001).

This can present a serious problem. Because they are essentially atheoretical (i.e. story-free), factor-based models cannot offer an *explanation* when a factor associated with risk in one context turns out to be associated with resilience in another, which is not uncommon[7]. Hence, on their own, they are poor guides for action.

More importantly, this kind of purely statistical approach finds it difficult to discriminate between those variables that 'matter' (causal factors) and those that don't (markers, symptoms, irrelevant statistical associations) for the purpose of prevention. This is a significant issue in a field like CP, where hundreds of factors which correlate with crime have been identified. Who is to be targeted? What interventions should be prioritised? Statistics cannot say.

Yet, to prevent a problem from occurring one must disrupt its causes, not just attack the symptoms associated with it. Breaking a barometer does absolutely nothing to disrupt the weather. In the search for an actionable knowledge-base, one needs more than a laundry list of statistical correlations, however rigorously produced[8].

One needs causal explanation – a narrative which sets out how and why one thing brings about another.

The argument for 'good stories'

Testament to the importance of this causal narrative is that a good story is one of the hallmarks of mature science. New theories are rarely accepted by the scientific community until a plausible causal process has been conjectured, which makes sense of observations.

As John Eck (2005:708) illustrates:

[E]arly proponents of continental drift were unable to persuade geologists that their theory of continental movement was valid, despite the considerable evidence they amassed. It was not until 1965 with the elaboration of the underlying mechanism [heating of the earth's mantle creates convection currents] (and evidence for that mechanism) that geology accepted the idea that the earth's crust moves […].

Bringing the discussion back to the matter of building confidence in OIAs, this indicates that an effective OIA evaluation framework should be one which produces explanations, as well as robust MoEs – one which addresses the challenge of generalisation, as well as attribution.

The realist approach to evaluation is discussed at length in the next section, with particular emphasis on the strategies employed to tackle the problem of generalisation.

---

[7] For example, the same factor, 'community cohesion', is associated with a heightened risk of involvement in political violence in the context of nationalist terrorism, but with a lessened risk in the context of home-grown terrorism.

[8] For further discussion of this point, see Wikström (2007, 2011).

Further reading

Sherman, L. (1998). *Evidence-Based Policing*. Washington, DC: Police Foundation. Available from: http://www.policefoundation.org/content/evidence-based-policing

Welsh, B.C. and D.P. Farrington (2006). *Evidence-Based Crime Prevention*. In B.C. Welsh and D.P. Farrington (eds.), *Preventing Crime: What Works for Children, Offenders, Victims and Places*. New York, NY: Springer.

Tilley, N. (2009). *Crime Prevention*. Cullompton: Willan.

Sampson, R. (2010). "Gold standard myths: Observations on the experimental turn in quantitative criminology." *Journal of quantitative criminology*, 25: 489-500.

SECTION 5

Realist evaluation: Wrestling with the problem of generalisation

Devil's Advocate *par excellence*

Realist evaluation (RE) is theory-driven. This statement has profound implications, not only for the design and conduct of evaluation activity, but for the design and conduct of the social activity (CP initiative, health improvement programme, IO) being evaluated.

From the realist perspective, the commissioners, designers and implementers of social change activities are engaged in a scientific endeavour, even if they are often unaware of it. What is an IO, the realists argue, if not the implementation of an idea (in other words, a theory) about the kind of activities (*treatment*) which can be introduced into a particular social environment (*study population*) to effect a change (*treatment outcome*)?

Evaluation, then, is the process of validating (or invalidating) the treatment assumptions implicit in an operation. As Pawson and Tilley (2004:2), the fathers of RE, put it:

"When one evaluates realistically, one always returns to the core theories about how a programme is supposed to work and then interrogates it – is that basic plan sound, plausible, durable, practical, and above all, valid?"

The realist evaluator is a Devil's Advocate *par excellence*, whose unrelenting advocacy is put to a specific purpose: to improve future operations. If the realist evaluator doesn't stop at establishing effectiveness (or measuring success), it's because she knows that evaluation serves a greater purpose: it provides the lessons – *the knowledge-base* – upon which future operations will be founded. Hers is a drive for constant improvement.

This requires more than cataloguing and measuring the operation's outcomes. It requires *making sense of them*.

Deconstructing operations

For the realist, operations are *"theories incarnate"* (Pawson & Tilley 2004:3). The theories behind an operation are often complex. Paradoxically, the more complex they are, the less likely it is that the designers of the operation have articulated them fully, or at all.

The first task of the realist evaluator is to uncover these assumptions (or 'black boxes'), as comprehensively as can be managed. This is usually achieved through analysis of the operation's documentation, as well as interviews with the operation's designers and implementers before the activity starts, and through careful collection of (qualitative and/or quantitative) observations once the activity is underway.

Uncovering all of the assumptions implicit in an operation can be quite an undertaking. It is not unheard of for a programme of activity to be rolled out without much thought being given beforehand as to the reasons 'why'.

Eliciting theories

Box 2 overleaf briefly sketches a prototypical IO, which relies on social media to propagate messages with the aim of changing the behaviour of a specific group of individuals. It then presents a list of the many assumptions which underpin such an IO. The list, though long, is incomplete. Many more assumptions could have been elicited.

Some of these assumptions may turn out to have been warranted, but not others. If one wants to learn from this particular operation in a way that will benefit as wide a scope of future IOs as possible, it is necessary to establish which of these assumptions were supported in the end, and which weren't.

In Box 2's imagined social media-enabled IO, the assumption that communicating with group members via social media platform can change the group's behaviour might turn out to have been valid. However, the assumption that the behavioural change would last might not have been verified (i.e. they resumed the undesirable behaviour in month 7, soon after the operation ended). Such a finding might suggest the following lesson: that social media cannot achieve lasting behavioural change unless it is sustained (for more than 6 months).

Cue the next social media-enabled IO, built along a similar design, which goes on for a year. Once again behavioural change fails to take. The lesson of that evaluation, which builds upon the previous, is that social media influence is unlikely to achieve lasting change. One possible explanation for this is that this sort of activity affects the situational (read: temporary) factors which impact behaviour (such as motivation or perception of the capability to carry out an action), instead of affecting dispositional (read: lasting) factors (such as propensity; i.e. the individual's moral filter). This new *'theory'* will inform the design of the next IO. And so on. In this manner, *each new evaluation strengthens the OIA knowledge-base.*

At the end of the day, the evaluator wants to identify the operation's *key ingredients*: those elements which are responsible for the IO's success (or its failure). Some of the assumptions which drive the IO will be more important than others. These core theories will be the focus of the realist evaluation. Time and resources are finite; therefore, less fundamental assumptions will be weeded out.

If repeated evaluations fail to validate the assumption that attitude change leads to behaviour change – a core assumption behind much influence practice – it might be time to revise the theoretical underpinnings of IAs. One would then look for inspiration to models of behavioural change that don't rely on attitude change as an active ingredient.

Box 2 Uncovering an operation's 'black boxes'

Picture this: *an online influence operation, which involves setting up an identity on a social media platform for the purpose of promoting the adoption of a particular viewpoint or attitude among the members of a specific group, in order to get the group in question to desist from a particular course of action. The operation is carried out over six months.*

Such an operation is bursting at the seams with theories, large and small. To uncover them, ask yourself:

"What are the assumptions (the conjectures or sometimes simple *guesses*) that have gone into the crafting of this operation?"

The designer appears to have assumed the following:

1. that attitude determines behaviour;
2. that changing a person's attitude is enough to change their behaviour;
3. that people's attitudes can be changed by an external influence;
4. that changing people's attitudes can be achieved through online interaction;
5. that a social media platform is an effective way to communicate in such a way as to effect attitude change;
6. that setting up a new identity on a social media platform is the best way to achieve the desired objective;
7. that setting up a new identity on a media platform is the best use of available resources in the quest to achieve the desired objective;
8. that a credible media identity can be set up in six months;
9. that six months is enough time to influence a group in a lasting way;
10. that the particular attitude or viewpoint being targeted for change is the cause of the group's undesirable behaviour in the first place;
11. that offline sources of influence will not counter the effectiveness of the online message;
12. that the group will not revert to its prior behaviour as soon as the operation ceases;
13. that any change in attitude will not result in unintended negative consequences, either among the target group or some unknown party (for example: the creation of the new identity spurs others on the same media platform to begin their own campaign of counter-influence);
14. that the social media platform will continue to operate for the next six months;

15. that the group members are rational agents, who behave in predictable ways;
16. that the group members will be culturally sensitive to the framing of the message propagated through the social media platform;
17. that the group members have regular internet access;
18. that they are susceptible to influence...

Think of all the ducks that need to be put in a row (all the core assumptions that need to be valid) for the stated objective to be achieved.

The list is not exhaustive. What appeared like a straightforward 'idea' at the outset turns out not to be so simple after all. Many of these assumptions will turn out to be ungrounded guesses. Some may never have been articulated by the operation designers.

Each is a 'black box', which the evaluator must open.

The earlier in the life of the operation, the better.

The quest for mechanisms

If theories are the broad narratives about the causes of change that drive an IO, mechanisms are the processes through which that change actually occurs.

As stated previously, in much of science causation is not generally assumed until a credible mechanism has been postulated (as in the example of continental drift)[9]. The postulation of a plausible mechanism is often what will prompt scientists to go from talking about an 'association' or correlation between two factors, to hypothesizing that one is the cause of the other. Interest in this mechanism-based view of causation has been gaining in many areas of social science, inspired by the state of affairs in the natural sciences.

Since evaluation is, in essence, an exercise in trying to assess the support for a causal relationship between the operation and its outcome, uncovering and testing for the presence of mechanisms is a foremost task for the realist evaluator.

On this point, two important remarks:

1. Mechanisms are generally unobservable. Gravity cannot be seen, only its effect. Mechanisms are inferred from data or deduced through logic from theory. The mechanism is not the same as the 'measure' being put in place. A variety of measures can activate the same mechanism, and one measure can activate more than one mechanism.

[9] For a full discussion of the role of mechanisms in science generally, see Bunge (2004).

2. Social causation, and the subsequent activating of social causal processes (social mechanisms), is almost inevitably the result of causal interaction. For example, a crime is the result of a perception-choice process (mechanism) which results from the interaction of an individual with a particular propensity for action and a situation with particular criminogenic features[10]. Single-cause explanations of social events are rare, if any exist at all.

Conjecturing a mechanism involves *making an educated guess about the link between the operation's key ingredients and its outcomes*. Without positing these mechanisms, the evaluator cannot generalise and draw lessons for future operations.

Examples of measures and mechanisms implicated in CP include:

- increasing the perceived effort of stealing a car (the mechanism) by giving away free steering locks with every car purchase (the measure);
- influencing perceptions of risk (the mechanism) by publicising that a policing operation aimed at cracking down on residential burglaries is under way, in the local paper and on the news (the measure);
- removing a perceived provocation that could provide motivation for disruptive behaviour (the mechanism) by offering halal meals for Muslim offenders in prison (the measure).

Operations and programmes generally involve several mechanisms, which may need to work in tandem to produce the desired outcome.

A rule of thumb: *if the operation designers cannot, at the outset, explain how an operation will achieve its objective – i.e. through which (plausible) mechanisms – it is unlikely to be successful.*

This basic challenge to any planned operation (*'Show me your mechanism'*) may end up sparing an organisation a lot of wasted time and effort.

Research designs and the anticipation of unintended effects

Eliciting key theories and positing mechanisms serves yet another purpose: it guides the evaluator in the elaboration of the research design and the all-important task of data collection.

One of the particularities of realist evaluations is that realist science is, as Robert Sampson puts it, *"catholic on method"*. It is not assumed that one type of research design (such as the randomised control trial) trumps all others when it comes to evaluating social change initiatives in natural settings. There is no 'gold standard'. Instead, *there are appropriate methods to answer specific questions.*

Some questions may necessitate qualitative approaches, other quantitative designs. Before deciding which method to choose (e.g. focus group; in-depth interviews; time series analysis)

[10] For a fuller discussion of the problem of causation, see Wikström (2011).

and what data to collect, one must know what questions are being asked. This is where the elicitation of theories comes in.

Eliciting theories from the operation's designers and commissioners – making explicit what is often implicit – is also the first step towards anticipating unintended consequences of the operations. *Action can provoke unintended reaction.*

Dixit sociologist Robert K. Merton (1936):

"[W]ith the complex interaction that constitute society, action ramifies. Its consequences are not restricted to the specific area in which they are intended to center and occur in interrelated fields ignored at the time of action."

Preventing burglary in one neighbourhood may, if certain contextual features are present, displace the problem to another. Anticipation means that the evaluator can select a research design and identify a class of data which will allow for the monitoring and detection of the unintended effect (here, taking measures of the problem in adjacent neighbourhoods).

Because it is impossible to implement all-encompassing research designs and to collect perfect data, anticipation is key.

The importance of 'context'

As previously stated, the realist approach is particularly sensitive to the role of context. This is the crux of the contention regarding the use of randomised control trials (RCTs) as an evaluation methodology.

The main purpose of RCTs is to rule out all sources of influence aside from the 'treatment.' For example, in a drug trial, the RCT would control for (rule out the impact of) such 'ingredients' as patients forgetting to take their medicine and others neglecting to fill in their prescriptions because they didn't have time to run to the pharmacy during business hours.

By contrast, the realist will look to detect and understand the role of these context-specific ingredients. Once the treatment is rolled out 'in real life with real people', the realist argues, these ingredients will play a part, so *they have to be accounted for*. A well-controlled drug experiment will tell you much about the efficacy of the treatment once introduced into a human *biological system*. It won't tell you much about its effectiveness once introduced into a human *social system*. Hence, evaluations must take into account the effect of the system into which intervention takes place, or fail to achieve critical understandings.

If one wants to understand how an IO will perform under 'real conditions', context-specific ingredients must be treated as part-and-parcel of the process of change, not ruled out of evaluations because they might 'pollute' the findings. Contextual ingredients – more specifically, the way they interact with the operation's own ingredients to produce outcomes – *are* the findings.

One of the most challenging tasks for the realist evaluator is, therefore, to *understand enough about the interaction between the characteristics of the operation (its measures and*

mechanisms) and the characteristics of the context in which it is implemented, to draw lessons about the conditions under which a future operation is likely to succeed – or likely to fail.

Mechanisms tend to be highly context-sensitive. Interaction with inauspicious contextual features may prevent a mechanism from being activated and achieving the expected outcome. (More on that in the next section.)

Operating in 'open systems'

The study of the interaction effects between the features of social interventions and the features of contexts is complicated by the fact that initiatives are implemented in *open systems*.

Characteristically, *open systems are subject to multiple sources of influence*. Several interventions may be taking place at once, either administered by the same organisation, or by rival outfits. Influence can be exerted by informal agents, such as the media, politicians, civil organisations, business, networks of acquaintances, friends and family.

Open systems are subject to the knock-on effects of changes taking place in other systems, such as large scale political and economic changes. These changes reverberate across levels of analysis (i.e. from macro to micro and back). Think of the impact of monetary policies adopted in Brussels, which, through the complex, often poorly understood knock-on interactions of economic systems, affect the decisions and behaviour of families in another part of the world.

Think, similarly, of the impact kinetic operations may have on IOs. In a RAND report on the effectiveness of PsyOps in Afghanistan, the author observes that IOs ran afoul of perceptions shaped by kinetic operations taking place concurrently – such as house searching, or, in extreme cases, airstrikes which caused civilian casualties.

No IO takes place in a vacuum. Case in point: Events which take place offline can have tremendous impact on events which take place online, and vice versa. Indeed, if events in one system could not affect events in another, there would be no point undertaking OIAs in the first place.

System permeability can be an advantage when harnessed and a drawback if ignored. Evaluation activity must establish what role system permeability played in the operation's outcome.

Understanding operation-context interactions

Guided by theory, it is the job of the realist evaluator to identify the external factors which impact operational outcome. Some may turn out to be key ingredients in success or failure. Others may turn out to be the real cause of the change being observed, meaning success or failure would have been wrongly attributed to the operation had this factor not been considered.

Picture the impact of the release of a new iPhone on snatch-and-grab figures in London. A police operation intended to reduce this type of crime, which coincided with the iPhone release, might appear to perform less effectively – having to deal with a sudden rise in opportunities for

theft and an increase in offender motivation – than the same operation implemented at another time.

The evaluator must root out such changes in the operation's environment, in order to control for their impact in her research design. Perhaps, when the sudden influx of new attractive smartphones on the streets is ruled out, the police operation is shown to be relatively successful after all. If it hadn't been implemented, the numbers would have been worse. The realist evaluator doesn't want to throw the baby out with the bathwater without good reason.

Equally, the realist evaluator wants to keep track of the way an operation might affect the conditions of its own future success. Intervention is change, and change can create risks and vulnerabilities where there were none. For example, publicising the successes of medical research in developing treatments for HIV, to the extent that the people infected can expect to live a near-normal life, might affect public perceptions of risk, leading to a decline in condom usage and an increase in cases of HIV.

In the case of Box 2's hypothetical operation, setting up a new social media identity might prompt others to create identities of their own in order to counter the perceived influence. This is an example of the well-known problem – in CP and other domains – of escalation. Keeping track of unintended system change is an important task of evaluations.

Realists are, by definition, pragmatists: they accept that *some element of self-defeating change, or a decrease in effectiveness over time, is inevitable*. That is why operations must be evaluated routinely for improvement.

Theory failure vs. implementation failure

On the subject of throwing babies out with the bathwater, realist evaluators recognise that the factors involved in the implementation of an operation – its delivery – play a crucial role in the final outcome of the activity.

Much as the evaluator wants to attribute correctly which effects are due to operational activity and which are due to contextual variations (as in the iPhone example), they also want to correctly attribute those effects which are due to implementation factors. Remember: the chief aim of the evaluation is to learn the right lessons from the operation, not just to measure effectiveness.

Consider once again the example in Box 2. Some of the assumptions elicited are clearly related to a particular theory of influence (e.g. 'it is possible to influence people's attitudes through online media'; 'six months is enough time to achieve this'), while others are related to implementation (e.g. 'the social media platform will continue to operate for the next six months'; 'the targeted group has regular internet access').

If all implementation assumptions are met (stable media platform, messages rolled out on schedule and according to plan, group access to the internet is confirmed), but the operation doesn't deliver, we may be dealing with a case of partial or total *theory failure*. We might conclude that some or all of the designers' theoretical assumptions – about the capacity of online

media to influence attitude, about the role of attitude in determining behaviour, and so on – were wrong.

If, however, it can be established that the operation was not delivered according to plan, then we may instead be dealing with a case of *implementation failure*.

Examples of implementation failure include:

- planning an operation which requires that 'treatment' be delivered by trained personnel, but finding out 'on the day' that such personnel are unavailable;
- counting on the cooperation of partners (e.g. civil organisations, foreign agencies), who, once the operation is under way, refuse to 'play ball';
- diffusing messages in a language the local population doesn't understand.

Implementation failure can occur at all levels of an organisation.

The aforementioned RAND report on PsyOps in Afghanistan catalogues a number of implementation issues at programme level, such as the long time between planning and execution due to delays in the approval process which requires going up the chain of command to battalion levels. As one can imagine, such delays could render communication measures useless in cases where response-time is critical.

An evaluation must clearly identify *which outcomes result from the failure (or success) of theory, and which result from the failure (of success) or implementation*.

It would not do to throw out good theory when it has never, in fact, been properly applied. Furthermore, lessons can be learned about factors which support effective implementation, and the need for implementation failure contingencies in future operations.

Evaluation comes in different flavours

There are several types of evaluations, which serve different purposes and are more or less relevant to the present remit of building confidence in OIAs.

In some cases, agencies want to carry out *impact evaluations* for punctual measures, knowing that the operation will not be repeated. The intent is only to show that, in this particular case, money has been well-spent or something has been achieved; it isn't to 'learn lessons' to be applied elsewhere.

Theory-of-Change evaluations are theory-driven, but rather than test the operation's underlying causal theories, they set out to test programme theories. Programme theories are theories about what is required to carry out a successful programme implementation (e.g. what kinds of resourced are needed; what kind of organisational structure and management style works best), as opposed to theories about mechanisms (e.g. theories of behaviour or theories of influence).

Theory-of-Change evaluations are most often used to evaluate complex programmes, which coordinate multiple initiatives or require rolling-out on a large scale[11].

Two other types of realist evaluations are of greater interest for OIAs: *formative* and *summative evaluations.*

Formative evaluations

Formative evaluations offer unique benefits in the early days of a new activity, which is why they should be prioritised in the first instance with OIAs.

They are also the most intrusive form of evaluation, to the extent that they require a high level of collaboration between evaluators, designers and implementers – before, during and after the operation is underway.

Formative evaluations are a species of *action research*, which involves the close collaboration of academics and practitioners within projects aiming to improve interventions in order to solve concrete social problems.

In the early stages of a formative evaluation, experts who have extensive knowledge of scientific theories and the scientific evidence-base in the relevant domain (here: theories of influence and behavioural research), work closely with operation designers to:

1. Analyse the problem and establish what is the objective of the operation (what change is being pursued);
2. Given 1), establish what is the theoretical basis for the operation (what relevant theories are out there; what are their respective evidence-base);
3. Ascertain what techniques are available to achieve the desired change in light of the theory or theories selected in 2);
4. Determine the delivery methods which will best enable the implementation of the techniques identified in 3);
5. Track the implementation of the operation and measure expected effects;
6. Make sense of (typically mixed) results and synthesise lessons learned.

The advantage of formative evaluations for operation designers is that *designers can avail themselves of expertise that (if the evaluator is well-selected) reflects the state-of-the-art in the science of human and social change*, without having to conduct onerous and often complex literature searches themselves.

For the academics, the benefit is obvious: they are granted a rare opportunity to test theoretical assumptions in real-life conditions.

[11] For a comparison between theory of change and realist evaluation, see Blamey and Mackenzie (2007).

Because these phases have an iterative character, the evaluators are on hand to help designers and implementers fine-tune the operation as it goes, rather than wait until after the facts to declare that something has failed. They are also most likely to detect counter-intuitive effects or unintended consequences of the activity, including unexpected improvements or side-benefits, allowing implementers to capitalise on these.

Formative evaluations require trust and the willingness to collaborate between individuals with different priorities and different stakes in the operation. For this reason, they can face practical obstacles (e.g. access to sensitive data: few academics have security clearance).

However, the rewards in terms of confidence-building are non-negligible. Evaluators can help designers formulate explicit predictions and convincing narratives of effectiveness (e.g. 'this technique achieved this [visible] change among this group of this population because it activated this [invisible] mechanism in this particular context').

Nothing builds confidence like an accurate prediction backed up by a plausible story of success.

Summative evaluations

The main difference between a formative evaluation, and a summative one, is that formative evaluators guide the formulation of theories and techniques underpinning the operation, rather than simply elicit them. In terms of Phase 5 and 6, however, the logic is the same.

In the realist tradition, summative evaluations have two components: *process evaluation* and *outcome evaluation*. This two-pronged approach owes to the need to correctly attribute responsibility for outcomes.

The process evaluation tracks the delivery of the operation: what was done, when, with what resources – on the ground, rather than on paper. It is through process evaluation that we can find out, for example, that 100 anti-burglary alarms were purchased, but only 50 were distributed to residents of the neighbourhood, 20 were installed, and only 10 were properly plugged-in and functional. A good process evaluation will even be able to tell you why only 10% of alarms on which money was spent were put to use in the end.

The outcome evaluation, meanwhile, monitors the intended and unintended consequences of the operation, in order to assess its effectiveness and the reasons behind it. Both evaluations can be conducted retrospectively, but the best process and outcome evaluations are planned from the start, concurrently with operation design or very soon after.

This owes to the need to collect antecedent data; i.e. establish what the state of things was *before* the operation was delivered in order to demonstrate that something changed *after*; and to establish in advance of time what kind of information must be collected as the operation goes on to properly test the elicited assumptions.

If run concurrently to the operation, process evaluation may even help address implementation failure as or before it occurs. In practice, process and outcome are almost never assessed separately, which is not a problem. As long as the evaluation includes measures of outputs (implementation targets; e.g. number of leaflets distributed) as well as outcomes

(ultimate targets; e.g. change in voting intentions), process patterns can be monitored and outcome patterns interpreted in light of that information.

Regardless of what it is called, the evaluation must deliver enough information to make sense of what happened. For IOs, the main lesson is this: when carrying out evaluations, *measures of effectiveness (MoEs) can never be enough*. MoEs provide the undeniably necessary picture of 'what worked where', but not 'why it worked there'.

MoEs do not an evaluation make. *Attention to theory and mechanisms is needed to construct a convincing and useful narrative of success.*

Improving upon realist evaluation

While it most certainly doesn't deny that attributing causation is one of the chief goals of evaluation, the realist approach unashamedly puts a premium on tackling the problem of generalisation, advocating that particular attention be paid to mechanism-context interactions. Given this emphasis on contextual effects, one would expect realist evaluators to have dedicated a great deal of attention to what is, in this report, referred to as *the problem of analysis*.

Put another way: one would expect that the realist tradition would have produced, if not actual tools, then a framework within which to analyse the features of open social systems, to help problem analysts and solution designers identify which aspects of the operation's context may interact with the operation's activities to affect the outcome.

However, at this juncture, context analysis seems to remain an act of imagination which rests almost entirely on the individual skills and expertise of individual evaluators. Yet, in the absence of a systematic approach, realist evaluation runs the risk of coming under fire, as it has in CP, for its lack of procedural clarity.

If evaluation is to become the standard in IAs, the problem of analysis must be addressed more rigorously. It will also be necessary to evaluate, or at least expand upon, the realist claim that operations are (and should inspire to be) chiefly 'incarnations of (scientific) theories'. As the case will be made later on in this report, operations are based on much more than theories about causes of change (e.g. mechanisms of influence). They also reflect rules of implementation and design, which come under the heading of engineering more than science.

As we will see, making the distinction between what belongs to the domain of science and what belongs to engineering and technology has more than academic consequences for the future of evaluation activities, as well as for the future of IAs. Reformulating the product of realist evaluations as *technological rules* instead of fuzzier notions would begin to address a chief criticism of realist evaluation – that it is more 'craft' than 'science' – while preserving its most valuable contribution: the grounds for generalisation provided by mechanism-based explanations, which highlight the imperative to go beyond mere MoEs.

But more on that later. In the next section, it is proposed that updating the realist logic with an analytical framework purposefully formulated to handle social systems is a promising way of attacking the neglected problem of analysis.

Further reading

Eck, J. (2005). "Evaluation for Lesson Learning." In N. Tilley (ed.), *Handbook of Crime Prevention and Community Safety*. London: Willan.

Pawson, R. and N. Tilley (2004).*Realist Evaluation*. Available from: http://www.communitymatters.com.au/RE_chapter.pdf

Pawson, R. (2006). *Evidence-Based Policy: A Realist Perspective*. London: Sage.

Pawson, R. and N. Tilley (1997). *Realistic Evaluation*. London: Sage.

Eck, J. (2006). "When is a bologna sandwich better than sex? A defence of small-n case study evaluation." *Journal of Experimental Criminology*, 2:345-362.

Astbury, B.and F.L. Leeuw (2010). "Unpacking black boxes: Mechanisms and theory building in evaluation." *American Journal of Evaluation*, 31(3): 363-381.

Bunge, Mario (2006). *Chasing reality: Strife over realism.* Toronto: University of Toronto Press.

Munoz, A. (2012). *U.S. Military Information Operations in Afghanistan: Effectiveness of Psychological Operations 2001–2010*. Santa Monica, CA: RAND, National Research Defense Institute. Available from: http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1060.pdf

SECTION 6

Going ‘systemist’: Dealing with the problem of analysis

From ‘context’ to ‘system’

To help decision-makers and operation designers figure out what activities might work to achieve their particular objective, the realist evaluator needs to establish what kind of measures activate which mechanisms in what sort of context to produce which outcome.

This process consists in identifying *context-mechanism-outcome (C-M-O) configurations*, which are the key transferrable product of a realist evaluation.

Isolating C-M-O configurations is easier said than done. Although ‘mechanism’ is a concept with a fairly long philosophical pedigree, ‘context’, like ‘environment’, is a much looser notion, and while identifying relevant contextual features in a small, well-contained initiative seems doable (e.g. hiring attendants to prevent thefts in a car park; see Table 1 overleaf), doing the same in light of a much more ambitious operation (e.g. conducting an online campaign to improve the perception of Coalition forces among national populations, in order to gain support for troupes in Afghanistan) is another kind of challenge entirely.

To date, the notion of ‘context’ remains imprecisely defined in the realist approach. While realists make great case of interventions taking place in ‘open systems’, they have not clearly defined what a ‘system’ is, nor how it should be handled concretely. Neither have they elaborated on the relationship between ‘system’ and ‘context’ – often seeming to use these notions interchangeably.

Table 1 Mechanism-Context Configurations for Car Park Theft

Measure	How it works (mechanism)	Works best if… (context)
Improving surveillance at deck and lot entrances/exits by improving lighting, removing obstructions and/or encouraging vendors to set up shop there	Increases thieves' perception of the risk of detection when entering and leaving the car park	…the facility's perimeter is secure (i.e. thieves can't get in any other way)
Hiring parking attendants	Improves surveillance of facilities, especially at entrances and exists	…the facility's perimeter is secure, so those who enter and exit must pass the attendant, and the attendant booth is designed to facilitate surveillance …the priority is reducing theft of cars, as this measure doesn't perform as well against theft from cars

Source: Adapted from Clarke (2002)

As a consequence, there is currently no realist methodology or analytical framework for identifying relevant aspects of context – those all-important features of the environment which are likely to interact with the operation's activity to produce intended and unintended outcomes.

Yet an evaluation framework destined for OIAs must be able to anticipate and harness issues of permeability, knock-on effects, and interaction effects – and do so reliably. If one is to capitalise on the strengths of the realist approach, notably the delivery of convincing and generalisable narratives, this analytical shortcoming must be addressed.

Thinking about systems systematically: The CESM model

Systemism, a philosophical system developed by physicist and philosopher Mario Bunge, offers some direction. Systemism is not in itself a theory, but an approach to research programme design, problem analysis, and, ultimately, solution design. Given that systemism belongs to the scientific realist tradition, realist evaluation's imprecise notion of 'context' can be easily subsumed under systemism's better-defined concept of 'system', without renouncing other valuable realist insights.

In systemism, a system is defined as *"a complex object whose parts or components are held together by bonds of some kind"* (Bunge 2004:188). Examples of systems are atoms and radar networks (physical systems), cells and horses (biological systems), business firms, terrorist groups, political parties, crime gangs, families, tribes, armies, and societies (social systems).

A system can be analysed in terms of the *CESM model*:

- *Composition* (C); the set of components that are part of the system
- *Environment* (E); the collection of items (including other systems) that act or are acted upon by the system
- *Structure* (S); the ensemble of relations (bonds or other links) that hold the system's components together (the endostructure), as well as tie the system's components to items in the environmental (the exostructure)
- *Mechanisms* (M); the processes that are characteristics of the operations of the system, some of which maintain the state of the system, others which effect change.

The CESM model of a system is a representation of that system at any given time, as each of these elements are subject to change.

The systemic approach differs from other forms of systems theory in that it doesn't assume that relations between the system's components is of a single, or main, kind (e.g. economic, political, cultural, biological, and so on). Rather, it considers that *all sorts of relational processes take place at once, and that none necessarily supersede the others*.

Nor does it assume that either structure (e.g. social bonds), environment (e.g. political institutions), components (e.g. actors) or mechanisms (e.g. psychological or biological mechanisms) are key to the explanation of social processes (including processes of influence), as do structuralist or rational choice approaches.

Instead, systemism considers that each element has its part to play, and *to neglect any of them is to operate based on incomplete information*.

Systemic interaction effects: Dealing with the 'embeddedness' of individuals in multiple social systems

Bunge (2004) uses the example of the nuclear family as unit of analysis to illustrate the CESM model. This example is expanded upon here to illuminate this analytical approach:

- The *components* of the family-system are the parents and their children. (In some cultures, it might include grand-parents, and sometimes more distant kin; as always, context rules.) This example shows well why the model of the system must be understood as a snapshot and cannot be held as static: Think of the impact an added component (e.g. a new baby) has on the behaviour of the family-system.

- The system's *environment* is made up of other systems and their components, which act upon, or are acted upon by, the family: the neighbourhood, the village, the tribe, the firms in which the parents work, the schools the children attend, the civil societies (political parties, cultural clubs, sports associations, online social networks, and so on) to which any of them belong, the army in which one of them happens to serve, the local government, and the more distant, national governmental apparatus.

  Given any sphere of activity (physical, biological, psychological, social, cultural, political, economic; each sphere a *meta-system*), the relative influence of these environmental items on the components of the targeted system will be ranked differently.

- The *structure* of the family-system is made up of the bonds that hold its members together; here, biological and psychological bonds, such as love, filial attachment or duty. It also includes external bonds, which link the system's components to the outside, such as bonds of kinship, friendship or trade.

- Finally, the *mechanisms* of the family are those associated with its essential functions. In most cultures, these would be mechanisms such as caring, nurturing, teaching and learning (the mechanisms which underpin the child-rearing process), as well as the sort of marital exchanges that typify relations between the spouses. When these mechanisms are disrupted or undermined, the system breaks down.

Systemic problems have systemic solutions

The systemist approach champions the view that systemic problems must be addressed *systemically*: one must pay attention to all the elements of the system, rather than intervene at a single level. The same rule applies if one wishes to implement any sort of system-wide change.

A common example of a failure to think systemically might be: providing humanitarian aid to address the needs of a population (a biological and psychological problem), but failing to address endemic corruption at the same time (an economic and political problem), so that most of the aid effort fails to reach its intended targets.

In particular, systemism cautions against the idea that one can effect change in the components of a system (e.g. individuals) without taking into account the system(s) into which they are embedded.

Consider the vast problem of preventing or reducing delinquency in adolescents. Because of the nature of the process of socialisation, which requires attachment (a bond) between socialiser and socialisee to operate, an intervention aimed at diminishing the delinquent propensity of adolescents (e.g. social skills and empathy-raising programmes delivered at school) must take into account the features of the social systems in which adolescents are embedded (see Box 3).

The intervention is doomed to failure if the targeted adolescents are strongly attached (*structure*) to peers and/or to parents who themselves have delinquent or criminal propensities (*composition*) and have the opportunity to pass on their crime-supporting views (i.e. 'teach'; *mechanism*) to the adolescents when they leave school (which is only one source of influence in their *environment*).

To stand a chance of success, the intervention must not only deliver the skills training, but carry out activities meant to increase the level of attachment kids feel towards their school, as well as address the antisocial tendencies of their unschooled peers and family members.

Let's carry this lesson over to the influence domain: Since all influence activities take place in social systems, it is necessary to identify which system the operation is to take place in (*operational system*), and which system (or component of system) is to be influenced (*target system*). Because of the permeability feature discussed earlier, operational and target systems may or may not be one and the same.

Either way, it will, as illustrated in Box 3, pay greatly to analyse those features of the system which are likely to interact with the IA in order to anticipate outcomes.

Likewise, if one counts on permeability effects by acting upon the components of one system in order to influence the components of another, it will pay to analyse the bonds which tie these systems together, to pinpoint plausible pathways of influence. If no such pathways can be identified, the hoped-for knock-on effects may fail to take place, or happen in unpredicted (and possibly destructive) ways.

Box 3 Why systemic analysis trumps factor-based approach

The systemic nature of social events explains why the 'risk factor' approach discussed in Section 4 is an imperfect guide for action. In several cross-sectional studies of delinquency, 'lack of parental attachment' is found to be a *risk factor* (a predictor) of delinquency. In other studies, however, the opposite outcome is found: 'lack of parental attachment' is a *protective factor* against delinquency. For many, this is a counter-intuitive finding. Why would feeling little love for your progenitors protect you against future criminality?

Looking at the composition of the system in which adolescents are embedded allows us to make sense of the discrepancy: parental attachment is a protective factor when the parents hold pro-social (crime-averse) values; it's a risk factor when the parents hold anti-social values or, quite often, are themselves criminals. Hence, in cases where one or both the child's parents have criminal propensities of their own, a lack of attachment to the progenitors will in fact protect the child from the criminogenic influence.

The transferrable lesson is this: *if one is to design an intervention which aims to change individual propensity for any given action, in a given direction, one should make sure that attachment to the new source of influence supersedes attachment to any other sources, which are pulling in the opposite (moral) direction (like families, schools, churches, gangs, social institutions of all stripes, or even role models in the popular culture).*

A common-sense example, perhaps, but so are many explanations in hindsight. The point is that *looking at individual-level factors alone does not provide an explanation* (and therefore an effective guide for action). Consideration of systemic processes provides a richer – plausible and *actionable* – picture.

Beyond 'Target Audience Analysis': From C-M-O to S-M-O

Systemic analysis, as advocated here, is several steps of sophistication beyond the kind of Target Audience Analysis (TAA) techniques employed in the marketing and political communication domains.

It provides a structured way of thinking about target-environment interactions, and, one step further, interactions between:

- features of the environment (CESM-related features of target and operational systems, including, but not limited to, characteristics of the population);

- active ingredients of operations (mechanisms and the measures that activate them), and;
- desired outcome.

The traditional C-M-O configuration of realist evaluation should, for the purpose of guiding operations in highly permeable systems (such as cyber environments) be upgraded to the following analytical product:

S (relevant aspects of the CESM configuration, including system-to-system relations) →

M (measure-to-mechanism, including cross-level mechanisms) →

O (Outcome, intended and unintended)

Once identified, these products will make up the basis of the concrete, sharable and transferrable products of OIA evaluations. As they accumulate and are translated into plausible causal narratives, confidence in OIAs will grow.

The challenge of policy transfer: Tackling the risk of systemic failure

In a paper on the limitations of RCT evaluation designs, criminologist Robert Sampson elaborates upon the importance of understanding the impact of system embeddedness for policy, highlighting that:

"once a policy takes effect the rules of the game change, possibly inducing system level changes" (Sampson 2010:494).

This, Sampson argues, takes us beyond the problem of generalisation. It is less about generalisation than it is about *transfer* from one level of analysis to another, because *"homology of processes [mechanisms] across levels cannot be assumed"* (ibid:495). In other words, what worked for a small group of individuals cannot be assumed to work once applied to a whole population.

Sampson takes as example a measure which involves taking black children who live in a segregated part of the city by bus to white schools with better resources, in order to improve their educational outcomes. Evaluations are carried out using RCTs, which provide evidence that the measure is effective. An ambitious policy is crafted to roll out the approach nationally and tackle the burning problem of education inequality.

Though presented as a thought-exercise, the example is not hypothetical. What happened in practice was that white families, who seemingly did not want their children schooled alongside 'too many' black children, chose to leave the targeted areas for new pastures. The end-

result was schools that were filled mainly with children from socio-ethnic minorities, in neighbourhoods with a decreased tax-base, and therefore fewer educational resources available.

Though the measure did well in experimental conditions, the effect of implementing the initiative *as a policy*, system-wide, was segregation. The white families' reaction to the policy of forced mixity rendered moot the lessons inferred from the evaluations. Once rolled out, the measure encountered systemic conditions which interacted with the intervention's 'ingredients' in wholly new and unpredicted (though not *de facto* unpredictable) ways.

When explanations inferred from evaluations fail to predict the outcome of system-wide policies, we may talk about *systemic failure*.

It's important to note that observational (realist) evaluations would not necessarily have performed better in the example of the busing initiative. At issue is *the kind of theoretical framework adopted, implicitly or explicitly, at the outset,* which drives the choice of models, methods and analytical techniques to be used in the design of research programmes, interventions, and, eventually, evaluations.

Those frameworks must *set out to investigate and articulate cross-level mechanisms*. Failing that, they cannot hope to anticipate the effect of operations once they are implemented in open social systems – where they will likely have to contend with, among other things, recalcitrant targets who refuse to behave in expected ways.

More on the characteristics of theoretical frameworks which can best support a systemic approach and help evaluators deal with the problem of analysis in Section 7.

Further reading

Bunge, M. (2004). "How Does It Work? The Search for Explanatory Mechanisms." *Philosophy of the Social Sciences*, 34(2): 182-210. Available from: http://www.gemas.fr/dphan/cosmagems/docs/socio/PhilosophyOfTheSocialSciences2004Symposium_2Bunge.pdf

Bunge, Mario (2006). *Chasing reality: Strife over realism.* Toronto: University of Toronto Press.

Bunge, M. (2006). "A systemic perspective on crime." In P-O Wikstrom and R. Sampson, *The explanation of crime: Context, mechanisms and development*. Cambridge: Cambridge University Press.

Sampson, R. (2010). "Gold standard myths: Observations on the experimental turn in quantitative criminology." *Journal of quantitative criminology* , 25: 489-500.

SECTION 7

Building a knowledge-base for operation and evaluation design

Good theories help you understand what the problem is in the first place (and what it will take to solve it)

The point has been made that influence operations can only be as good as the theories that drive them. It is perhaps less evident that theories, like operations, can and should be evaluated for fitness.

To begin with, a good theory speaks directly to the matter at hand. A statement of the obvious, perhaps, but which drives home an important point: the choice of the theoretical approach which guides an IO is dictated by the *problem analysis* and subsequent *problem statement*, which motivated the IO in the first place.

What may seem straightforward ("In order to achieve our strategic goal G, We want population P to adopt attitude A so they will be more likely to perform behaviour B") is in fact not so. This statement is already weighted with theoretical assumptions of the kind an evaluator should elicit and assess, such as 'attitudes can be induced' and 'attitudes shape behaviour'.

A good knowledge-base will inform the formulation of the problem statement, a process which is most evident when conducting a formative evaluation. That first step is crucial, because *a problem wrongly formulated cannot be solved*. The experience of CP suggests that crime reduction efforts are often wasted or largely ineffective because the problem is badly-stated, and thus the objective of the intervention is wrongly identified at the start.

Let us brainstorm our hypothetical operation to achieve goal G:

For one thing, behaviour B may just be a symptom of the problem, not its cause, and altering it will do nothing to achieve goal G. Even if behaviour B is indeed a cause, attitude A may be wrongly assumed to determine behaviour B. The criminological knowledge-base suggests that no reliable, causal [attitude → behaviour] mechanism has been found, and that influencing population P's perception of Situation S by altering some features in their environment is a more reliable mechanism to effect a change in behaviour B.

But how long-lasting and pervasive does this change need to be? Changing behaviour can be achieved in the short-term by altering situational features, assuming that we have a valid theory of situational action. A long-term change, however, requires either a permanent, or at least sustained, change of the features of Situation S, or a change in population P's propensity for behaviour B. For this, we need a valid theory of propensity development.

This takes us into the realm of socialisation (effecting lasting changes in people's minds) as opposed to conditioning (effecting short-to-medium term changes in people's behaviour). Socialisation requires long-term commitment, better coordination, and access to much greater resources than is needed to achieve a short-term change in behaviour B.

Let's refine our example: If goal G is to de-radicalise a population P (i.e. remove their propensity to perceive terrorism as a plausible action-alternative, which is a long-term change), then a sustained, developmental approach is required. If goal G is to reduce the number of terrorist attacks committed by population P with short-term effect (a behavioural change), then a situational approach is appropriate (e.g. inducing perceptions that terrorist attacks are risky, unrewarding, or require higher capability than population P can muster).

Short-term and long-term change rely on different mechanisms, and the difference in the amount of investment required to effect this change is equal to the psychological distance between compliance on the one side and conversion on the other. One can be achieved with minimum involvement, but is temporary; the other is lasting, but will demand a more committed approach.

So what, exactly, is it that we should be aiming to achieve, given our overall objective?

Armed with this sort of analytical exercise, supported by the knowledge-base on behavioural change, it is possible to define goal G explicitly and confirm that it is, indeed, the goal we want to achieve, establish whether changing attitude A is truly the best way to achieve it, as well as acquire a sense of the effort which will be required to attain goal G. All this before any part of the operation has been implemented.

We may ultimately choose to reconsider going ahead with the operation altogether. In the final analysis, that is a question for planners, not operation designers, but now it is informed by a clearer statement and analysis of the problem.

A good theory will help you determine *whether it is better to do nothing, than to do something just for the sake of it.*

Caveat: the knowledge-base guiding this type of analytical exercise must be *commensurate to the problem-space*. If the objective of IOs is to change behaviour, then the knowledge-base

should rest upon scientific models of behavioural change (as opposed, for example, to models of attitude change, unless these models clearly articulate the mechanisms linking attitude to behaviour, rather than merely assuming them).

Hence, if the objective is to sway individuals who already hold strong beliefs, as opposed to unformed opinions, knowledge-bases located in commercial, social, political or even public health marketing should be approached with caution. There is a *difference of magnitude* between using soft techniques to 'influence' someone towards the purchase of a pair of jeans when they intended to buy a pair of jeans all along; to 'nudge' someone towards choosing energy-saving light bulbs instead of regular when they have no opinion either way; or to 'sway' the undecided in an election in the absence of any meaningful stakes for the 'influencee'; versus changing the mind or behaviour of an individual already committed to a cause.

Choosing an inappropriate framework may result in a waste of time, or worse. For example, empirical research in fields such as moral and political psychology suggests that improperly attempting to influence individuals already committed to a particular viewpoint can entrench them even more into their beliefs (in common parlance, it can 'radicalise' them).

Major General Andrew McKay and colleagues (2012) address precisely this issue in their critique of RAND's evaluation of U.S. Information Operations in Afghanistan, when they question *"the folly of attitudinal communication*."

They state:

"RAND has missed THE fundamental failing in not just US IO and MISO/PsyOps but wider ISAF efforts as well: A naive and immature understanding of the very process of communication in non-compliant conflict environments and misplaced confidence, and over reliance, upon marketing and advertising principles.

[We advocate] that marketing and advertising must now be considered as an utterly failed model for IO and MISO/PsyOps, one which must now be discarded in favour of a behaviorally-led approach embracing proper, proven, social and behavioural science" (emphasis as original).

A reading of the scientific literature informed by a more appropriate statement of the problem, they add, demonstrates that the relationship goes *change in behaviour → change in attitude*, and not the other way around.

If that is the case, the implications for IAs and OIAs are profound.

Good theories are mechanismic

If the main product of evaluations is an understanding of how particular measures activate certain mechanisms in a given system to achieve a specific outcome, then the theories driving the operations must be mechanismic[12].

Said one way: Theories or models should *explicitly refer to the causal mechanisms involved in producing the outcome, rather than just describe the factors associated with the outcome* (in other words, they should do more than list factors that correlate with the outcome, such as factors of 'vulnerability' or 'resilience' to influence).

Said another way: They should, as much as possible, *not contain 'black boxes'.*

Conjecturing mechanisms is the highest level of explanation a theory can achieve. Of course, these mechanisms need to be plausible and compatible with established scientific laws and observations. The quest for theory-refinement is a quest for ever-deepening explanations, which conjecture always more concrete (i.e. material) mechanisms.

To illustrate, in the study of criminality, observations that serious criminals tend to thoughtlessness led to the formulation of impulsivity as a personality trait associated with criminal behaviour. A general theory, which states that low self-control is the main individual determinant of crime (and the underlying factor behind impulsivity), eventually rose to prominence[13]. Self-control was then recognised as one aspect of what are now referred to as executive functions, a group of brain functions which sit in the pre-frontal cortex, the area of the brain involved in self-regulation and decision-making.

Identifying the underlying neurological mechanisms of self-control allowed for experimental, cross-sectional and longitudinal study, which is right now deepening our understanding of how individual self-control is established and maintained, what role it plays in behaviour, how it interacts with other (e.g. affective) brain systems, and under which social circumstances it fails or, conversely, is shored up.

Down the line, *the availability of deeper explanations means more control over the problem.* In this case, it opens up a slew of possibilities, from increasing self-control through childhood intervention to designing out environments that lead to its depletion.

Good theories are interdisciplinary, falsifiable, and no simpler than they need to be: The counter-example of rational choice models

In the social sciences, deep explanations tend to be characterised by their *interdisciplinarity*. They emerge at the intersection of biology, social cognitive neuroscience, psychology, sociology,

[12] The term is borrowed from Mario Bunge (2004) to encompass all forms of mechanism-based explanations, not just mechanistic (i.e. mechanical) ones.
[13] For the seminal statement of self-control theory, see Gottfredson and Hirschi (1990).

social ecology, statistics, and so on, and integrate and make sense of findings across cognate knowledge domains[14].

One of several objections to the usefulness of rational choice models (RCMs) in social science is that they make no reference to deep mechanisms and largely fail to take into account evidence showing that human behaviour and decision-making is underpinned by dual systems (cognitive *and* affective), which highlight the role of automaticism, cognitive biases, and other irrational mechanisms in the production of human judgement, human decision-making, and, ultimately, human behaviour[15]. Because they do not take this established knowledge into account, RCMs are judged to be unrealistic (therefore, *un-realist*).

Another objection, just as crucial, is that the key 'process' conjectured by most RCMs – that individuals seek to maximise their subjective utility – is a black box that can never be opened, since subjective utility cannot be measured. This means that demonstrating that a given behaviour *isn't* the outcome of utility maximisation, and that humans *do not*, in fact, *always* seek to maximise their utility, whatever the circumstance, is impossible to show. All of human behaviour can be 'reinterpreted' as being in the actor's self-interest *from her perspective*[16]. In other words, the theory can be made to fit any set of events.

'Unopenable' black boxes make theories unfalsifiable. Indeed, Becker (cited in Bunge 1996:374), claims that since *"rationality can be pretty flexible and the data are often limited, I don't frequently encounter decisive evidence against rationality"*, hereby sparing RCMs the risk of refutation. This is a cardinal sin in science, and furthermore a serious problem in the context of an evaluation.

If a black box cannot be opened, then an explanation of what happened cannot be produced. If a theory cannot be falsified (i.e. proven wrong), then *why bother conducting an evaluation at all?*

Some theories are attractive to academics and practitioners alike, because of their parsimony (they are relatively simple, compared to others). But a full statement of Occam's Razor reminds us that the point is not to choose the simplest theory by default; it is to choose the simplest between two or more theories *which all explain the problem equally well* and are *equally-well supported by the evidence*.

Shorter: In the matter of theory selection, *we should go for the optimum level of complexity, not the maximum level of simplicity*.

---

[14] This is why any given problem space should avoid becoming the *chasse guardée* of a single discipline – a problem known as 'disciplinary capture'.

[15] For an accessible, engaging and seminal text on this topic, see Economics Nobel prize-winner Daniel Kahneman's (2011) *Thinking Fast and Slow*.

[16] For an extended discussion of this problem, see Hodgson (2012).

Good theories are general

Mechanismic theories, so long as they refer to concrete mechanisms, will have the benefit of general application within the boundaries of their problem domains. Science aspires to generality: it wants to explain many cases; not one or a few. By contrast, problems are always local: they are the product of historical circumstances, which may never be repeated.

How, then, can we understand the relevance of general science to local problems?

Writing about the role of the theoretical knowledge-base in CP, Per-Olof Wikström (2007:72, 75) puts it this way:

"Local [CP] partnerships face very different realities. The problem profiles vary considerably. [...] However, the fact that the problems are different does not mean that the underlying causes of particular problems are different. I submit that the causes are the same, while the problems are not. The reason why partnerships face different problems is simply that the factors causing various problems differ among localities in their presence and strength. [...]

A well-developed and knowledge-based strategy (founded on an empirically-grounded theory of crime causation) would make it possible for policy-makers and practitioners to better focus their attention on the social, developmental and situational processes in which intervention can make the greatest impact in preventing or reducing crime and disorder."

Faced with a uniquely local problem, a well-supported, mechanismic theory will help operation designers *reduce seemingly unique, intractable complexity to a set of essential observations*. It will tell them where to look.

This is good news. Without general theories, we would have to start from scratch with every new problem.

Good theories are systemic (or compatible with a systemic approach) and expand the scope for action

To assist planners and designers with the strategic, analytical and practical challenge of carrying out operations in open social systems, theories must be systemic in their outlook.

This is not at all to say that the only useful theories or models are those which tackle system-wide events. Rather, it is to say that *a useful theory is one that plays well with other established theories and ideas, adding something to our understanding of causal processes at different levels of analysis* (individual, ecological, macro-social, and so on). It is a theory that enriches the playing field and, for the sake of a maturing scientific knowledge-base, is fertile in new, testable hypotheses and plausible conjectures about causes and causal mechanisms.

Here again, the experience of CP is instructive.

Following disenchantment with offender treatment programmes, which not only failed to significantly reduce reoffending, but, more importantly, didn't seem to make a dent in the crime rate, a group of scholars at the UK Policing Research Unit made the case that the crime problem could not be solved because it was *badly stated*. The ultimate goal of crime prevention was, in matter of fact, to prevent crime (an event), not criminality (an individual disposition). Hence, one should worry about crime and stop worrying so much about criminals.

Breaking down the factors involved in the emergence of crime events, they posited the crime triangle: for a crime to happen, a motivated offender and a vulnerable target (person or object) need to come together at a time and in a place, in the absence of a capable guardian who might deter the offender and/or protect the target. Hence, the purpose of CP should be to prevent this triangle from forming.

Until then, efforts had been aimed mainly at preventing the emergence of a criminal disposition in individuals. For this approach to be successful, proponents of the crime triangle argued, fundamental research on criminal propensity would have to yield a more robust knowledge-base than was currently available. Furthermore, this kind of intervention was resource-intensive and long-term; its effectiveness (or ineffectiveness) would not be measurable for some time. Finally, a focus on 'criminality' seemed to imply that only a special class of individuals, the 'criminals', were responsible for crime, when in fact most people had broken the law at one time or another.

The perception of rewards and risks (the perception of *criminal opportunity*) was hypothesised as the main causal mechanism of crime events, and the characteristics of places themselves, inasmuch as they shaped this perception, were said to play a causal role in the emergence of crime.

A whole new level of situational and ecological analysis was unlocked. From there, a 'criminology of place' was born, drawing from urban design, ecology, economics, management, administration science, architecture and computer modelling, to name a few.

This theoretical reformulation *opened up the field of possible interventions* in significant ways.

Why not, indeed, concentrate efforts on the immediate causes of crime, in order to achieve short-term reduction in the number of crime events? 'Hotspot' policing experienced a meteoric rise as the management and control of places – no longer just people – became the legitimate focus of policing activity. Marketing technologies, such as public messaging, were put to work alongside more 'kinetic' interventions, not as a tool to change offenders' beliefs or values (a long-term goal), but as a tool to influence their perception of the criminal opportunity *in the moment*, hereby influencing their decision to proceed, and, ultimately, their actions.

In their original formulation, theories of situational prevention relied on rational choice postulates to model offender decision-making. More recently, it has been recognised that this approach lacks realism, prompting a drive to reformulate perception-choice models, while holding onto the practical gains accrued by opportunity-based approaches.

Today the challenge is to integrate the knowledge-base on offenders and their propensity (developmental models) and the situational processes which give rise to their criminal behaviour in particular places at particular times (action models).

The point has been made that a rational and effective crime prevention strategy needs unified, systemic models, which integrate all the causes of crime, from proximal (immediate) to distal (ultimate), in order to guide coherent programmes, as opposed to 'patchy' or 'ad hoc' interventions based on whatever approach happens to be the 'flavour of the month'.

Ultimately, the choice of the overarching goal of any strategy (short- or long-term; temporary or lasting effects; cheap or costly; preventing the making of crime or the making of criminals; investing to tackle both sides) is a policy decision, which the scientific knowledge-base can only do its best to inform.

An ambitious policy agenda for IAs needs *a robust, well-integrated knowledge-base.* In its absence, it will be very difficult to produce meaningful problem statements and identify achievable objectives, or to plan and execute IOs in open social systems with any degree of control. Without a knowledge-base that is fit-for-purpose, IAs risk being irrelevant and ineffective at best, or counterproductive and damaging at worst.

Further reading

Wikström, P-O. (2007). "Doing Without Knowing: Common Pitfalls in Crime Prevention." In Farrell, G., Bowers, K., Johnson, S. and M. Townsley (eds.), *Imagination for Crime Prevention: Essays in Honour of Ken Pease*. Crime Prevention Studies Vol. 21.

Wikström, P-O. (2011). *"Does Everything Matter? Addressing the Problem of Causation and Explanation in the Study of Crime."* In McGloin. J.M., Sullivan, C. J. and L.W. Kennedy (eds), *When Crime Appears: The Role of Emergence* London. Routledge.

Felson, M. and R.V. Clarke (1998). *Opportunity makes the thief: Practical theory for crime prevention*. Police Research Series Paper 98. London: Home Office Research Development and Statistics (RDS). Available from: http://www.skywallnet.com/data_server/CA/OMT_PP_CP.pdf.

Bowers, K. and S. Johnson (2005). "Using Publicity for Preventive Purposes." In Nick Tilley (ed.) *Handbook of Crime Prevention: Theory, Policy and Practice*. London: Willan

McKay, A., Tatham, S. and L. Rowland (2012). *The Effectiveness of US Military Information Operations in Afghanistan 2001-2010: Why RAND missed the point.* Central Asia Series. UK Defence Academy. Available from: http://www.da.mod.uk/publications/library/central-asian-series/20121214_Whyrandmissedthepoint_U_1202a.pdf.

SECTION 8

Evaluation design: Balancing attribution and generalisation

"Data never 'speak for themselves' – making sense of causal patterns requires theoretical claims about unobserved mechanisms and social processes no matter what the experiment or statistical method employed [...]

The choice of method depends on the theoretical question and the nature of the phenomena under study, neither of which fall on a hierarchy. The hard truth is that ***we have little choice but to adapt in creative ways to the limitations that confront all social science inquiry****."*

Sampson, 2010

(emphasis added)

One evaluation, one design

When planning an evaluation, the choice of evaluation design is dictated by the questions the evaluation sets out to answer, and guided by the theoretical framework and problem statement which underpin the operation.

Good evaluation designs are *tailored* to the operations they assess. At the end of the day, no single method will systematically provide the 'right' answer, though some will always argue for or against their preferred approach.

Since this report aims only to provide a foundation for the evaluation of OIAs, this section will merely summarise principles of evaluation designs in terms of their relevance to the challenges identified in the introduction and expounded throughout: *attribution, generalisation, analysis and usability*.

Previous sections introduced the reader to the logic of evaluation activity. Designs are the concrete tools through which these principles are put into practice. Some designs are closely identified with one evaluation approach

(for example, evidence-based evaluations tend to be equated with RCTs); others are employed across approaches.

Like all tools, evaluation designs vary in their degree of sophistication. Some demand significant training and experience to operate safely. All of them set out to deal with the challenges outlined in this report and each of them are, invariably, better at dealing with some of these problems than others.

'What did we do?' Measuring the impact of activity

If nothing else, an evaluation has to establish what impact the operation had on the field of activity. This involves selecting or designing measures or metrics. Most often measures are quantitative (e.g. the number of people who registered on a forum), but they can also be qualitative (e.g. the content of pictures posted on a social media network).

Measures have to be selected as early as possible in the evaluation process. This is to ensure that measures can be taken before, as well as after, the operation starts. Gathering valid retrospective measures is hard to do and often constrains the evaluator to use measures which are not ideal.

Measures are valid to the extent that they actually represent what is being measured. This unwieldy notion is known as *construct validity*.

Put colloquially, *your measure is valid only 'if it means what you think it means'*.

As one can imagine, in a social world construct validity is often a lot harder to achieve than in the natural world. What we think of as 'data' comes bundled with all kinds of assumptions.

Think of what might seem like a straightforward measure, such as official statistics of recorded crime. Do crime statistics reflect the amount of crime that takes place, or do they reflect the decisions that police make about the crimes that should be recorded? What does an increase in recorded crime tell us? That more crime incidents took place, or that, for whatever reason, more victims chose to report incidents this quarter? More data are needed to figure this out.

Now consider the problem of measuring individual attitudes. We need some conceptual definition of what an attitude is. We then need an operational (measurable) definition. We need a tool to operationalise the definition; for example, a survey, which has to 'translate' with fidelity our operational definition into a series of questions. We need to demonstrate that the questions we have chosen actually capture our initial concept of 'attitude'.

We have to consider carefully the conditions in which the survey is administered. Perhaps we have captured something else, such as the desire to please the administrator of the survey (by providing the answers the respondent thinks the survey administrator wants to hear). Perhaps our questions were too suggestive of what we expected the response to be. Perhaps our initial conceptualisation was erroneous and what we have measured is something other than 'attitude' entirely. Perhaps a survey wasn't even the right tool. A growing number of social researchers are coming to question whether surveys tell us much of any use at all. Many

criminologists do not use them, preferring tools such as psychometric tests or structured interviews.

A rule of thumb: *the more 'remote' the measure is from the phenomenon we are actually interested in, the less valid it is likely to be.* (Think of how inaccurate and embellished a story can get the further removed it is from the original storyteller.)

In some circumstances, however, one may have no choice but to rely on '*proxy indicators*'.

Geocoding Google users' searches for 'flu symptoms' in order to track the spread of a seasonal flu epidemic is one example of using a proxy indicator. A cleverly designed proxy can spare a lot of effort. In this example, the alternative would be to call all general practices and hospitals to gather daily estimates in order to follow the progress of the epidemic in real time. This would be quite the undertaking. We could also track how many doses of the flu vaccine are being ordered and where they are being shipped, but remember our crime statistics example: vaccine orders may reflect policy decisions more than they do epidemic progression. Think back to all the vaccines that went unused during the swine flu pandemic of 2009. If a clever proxy is available, the loss of information may be worth it.

Devising valid measures is the main challenge of impact measurement. What will constitute a valid measure is wholly dependent on how the problem has been defined (e.g. are we trying to influence 'attitudes' or are we trying to influence 'behaviour' – the first may be a poor proxy for the second, much like citizens' beliefs about the amount crime that takes place are a poor proxy for actual crime rates).

Because the choice of measure will guide the choice of data collection method (e.g. carrying out surveys, conducting interviews, harvesting Google analytics), devising measures early allows us to plan for the evaluation activity as soon as possible. Linking measure selection to problem analysis means that measures can be designed to keep track of unintended or undesirable consequences. Part of that process will involve drawing a line beyond which effects will not be monitored. Total monitoring of operational effects in open systems is, of course, unachievable.

The chosen measures should remain the same throughout the evaluation, to rule out the possibility that any change uncovered is due to the change in measurement techniques, instead of a real effect.

Unfortunately, there is no sure-fire way of devising valid measures, which have to balance practical considerations with a ruthless questioning of assumptions. Trade-offs are inevitable, but should be scrupulously justified and documented. Data shouldn't be collected 'for the sake of it', without any idea of what it has to say. 'Data' are not the same as 'measures', much like 'information' isn't intelligence until it has been analysed.

Finally, if official data is used, evaluators need to be familiar with the protocols through which it was collected. This includes the use of survey data, such as reports from the Pew Global Attitudes Project.

‘How did we do it?’ Establishing accountability

Keeping track of what an operation actually entailed is a neglected component of evaluations, yet it is a critical part of the evaluation package. Evaluators must hold a faithful record of the operation’s outputs: *who did what, how and when*?

If an operation involved putting out messages over Twitter, how many went out, when were they posted, and who wrote them? If ads were run in the paper, when were they actually run, on what page, next to which article? What do we know about the paper’s circulation and the profile of its readership? Why did we pick this newspaper in particular? If the plan was to partner with a civil organisation, did they actually contribute? If not, why?

All of these details are needed to contextualise the evaluation’s findings; to come up with an explanation why particular elements of the operation have failed or succeeded (e.g. lack of cultural awareness, failure of technology, enthusiasm of the local commander for the project, unexpected support from local authorities); and how one might do better (or as well) next time. It may have been a great idea, but poorly executed. Perhaps the budget was insufficient, or conditions on the ground changed unexpectedly, and the operation had to be terminated too soon, despite the fact that it was beginning to show promising results.

It is not possible to rely on the original plans for the operation to answer these questions. This is, of course, because the saying ‘no battle plan survives contact with the enemy’ also applies to influence operations, though one might rephrase it as:

No plan for an influence operation can survive implementation in an open social system.

‘Did we do it?’ Attributing responsibility for change

The next order of business is to establish whether or not the operation’s activities caused any or all of the fluctuations revealed by measurements. Assuming the problem has been reduced, we want to find out whether the operation can take the credit. This means meeting criteria for causality and ruling out threats to causal attribution.

To make the case that the changes (i.e. outcomes) observed are attributable to the operation, it must be demonstrated that:

- *the operation’s activities are associated with the change(s).* <u>This is a matter of *statistical association*.</u> Statistical analysis of the variation between output measures (measures of operational activities) and outcome measures (measures of change) must show that both sets of measures correlate.

  Association, however, is not enough to demonstrate causality, because the relationship could, *theoretically*, go both ways. The outcome could be causing the output. (This seems counter-intuitive, but consider the relationship between crime and police activity. Statistically, it might look like an increase of police activity is causing a rise in crime rates. Causally, a *plausible* explanation is that as the crime rates rise, the police respond by doing more law enforcement.)

Output measures should include measures of *intensity*[17]. More activity should be associated with more change to strengthen the case that variation in the output is responsible for variation in the outcome.

- *the operation's activity precedes the change in the outcome measures.* This is a matter of *temporal ordering.* Causes always precede their effects. To add to the case for causal attribution, it is necessary to demonstrate that the outcome didn't change before the operation even started. This is why taking measurements *before* implementation is absolutely necessary. This is also why it is necessary to plan the evaluation at the outset, while the operation is still being designed.

- *the changes weren't caused by something else.* This is a matter of ruling out *rival causes.* This chiefly involves putting in place *controls*, so that one can estimate what would have happened had the operation not taken place (in scientific terms, we need to establish the *counterfactual*) and compare it against what did take place once the operation was implemented. It involves ruling out all sorts of threats to the integrity of the evaluation's design.

  This is quite possibly the hardest thing for any evaluation to achieve. It requires the evaluators to come up with a whole list of other factors, which could have been responsible for the change, and rule them out one by one, which is rarely attempted in practice.

- *the fact that the operation's activities is responsible for the change is the most plausible out of all other possibilities, given the points above.* This is a matter of establishing *plausible mechanisms* which could have caused the activities to effect a change in the outcome.

[17] See Bowers et al. (2004) for a technical discussion of intensity measures.

Table 2 Threats to internal validity

Threat	Explanation
History	The effect is caused by something that would have happened anyway, even if the operation hadn't taken place (e.g. due to 'seasonal effects'), or by some other event taking place at the same time as the operation (e.g. a counter-influence effort).
Selection	The effect reflects pre-existing differences between experimental and control group, or the fact that the target population was special in some way and particularly susceptible to influence.
Maturation	The target population had started to change anyway because of normal processes (e.g. the process of 'growing up') and would have continued to do so, even without the operation.
Measurement	The choice of measures, or the act of measurement itself, is responsible for the change (e.g. the same construct was captured using different measures before and after the operation).
Statistical regression	Also known as regression to the mean. The operation was implemented because the problem or issue was particularly bad. Things that are particularly high or low tend to naturally return closer to a normal state. In other words, even without the operation, an increase (or decrease) in the outcome measure would have taken place because of normal fluctuation.
Attrition	Also known as mortality. Evaluations often start with samples of a certain size, but participants drop out for varying reasons along the way. Attrition is a threat if those who drop out are different from those who stay the course, in a way that explains the effect (e.g. those who stayed were committed to change).
Direction of causation	It is not possible to establish whether an operational output is causing an outcome, or vice-versa.
Diffusion	Populations, groups, or individuals that were not targeted are somehow influence by the operation anyway (e.g. through the operation of social networks). This is a problem if they are part of the control population.

Source: Adapted from Tilley (2009) and Welsh and Farrington (2006).

There exists any number of threats to causal attribution. Some are threats to *statistical validity*, which undermine the search for a statistical relationship between outputs and outcomes; this can be because the sample of cases to evaluate is too small, or because an improper statistical technique has been used.

Others are threats to *internal validity*, which undermine the case for attribution of causal responsibility to the operation. Common threats to internal validity are listed in Table 2.

'What did we learn?' Providing grounds for generalisation

If all we wanted was to show that the operation had an impact, we might stop there, but the purpose of an evaluation is to provide a knowledge-base upon which to improve future activity. Above all, planners want to know: *'If it worked this time, will it work again?'*

Different approaches to evaluation have different strategies for dealing with this problem, as previously discussed. There are, broadly speaking, two ways to tackle it:

1. One can conduct many evaluations of the same operation implemented in different contexts (minimising as much as possible the variation in outputs between implementations), then subjecting the findings to a systematic review and statistical meta-analysis[18]. This is the strategy advocated by the evidence-based tradition.

2. For each operation evaluated, one can conjecture associations between mechanisms and contextual features responsible for the outcomes observed, based on analysis informed by the scientific knowledge-base, as well as the evidence generated by the impact and process components of the evaluation; if the next operation, informed by these conjectures, performs well, this is a test of generalisability. This is the strategy advocated by the realist evaluation tradition, as well as the systemist approach put forward in this report.

[18] See Welsh and Farrington (2006) for further discussion.

Table 3 Threats to external validity

Threat	Explanation
Setting attributes	The settings in which people develop and behave have characteristics which can vary significantly from one environment to the next. These characteristics may play an important role in allowing the effect of the operation to come about.
Target attributes	The characteristics of the populations, groups or individuals can vary significantly from one environment to another. These patterns of variation in target characteristics may be important in terms of the effect which was brought about.
Systems attribute	The features of the systems (composition, environment, structure, mechanism) in which target populations, groups or individuals are embedded will differ from one theatre to another. These differences may have a crucial role to play in the achievement of the objective.
Implementer attributes	The nature and characteristics of the people (commanders, analysts, front-line staff, agency) who implement the operation can vary from one operation to the next. These characteristics may be important in relation to the effect achieved.
Partner attributes	The characteristics of mediators or other partners (e.g. local authorities; media platforms) who actively assist the implementers in delivering the operation's activities will also vary. These variations may also play a part in the eventual outcome.
Dosage	The intensity with which the operation's activities are implemented and delivered differs between target populations or target settings. These variations in intensity may be important in explaining the effect achieved.

Source: Adapted from Tilley (2009)

Whatever strategy is employed, there are any numbers of threats to *external validity* (generalisability). The main threats are listed in Table 3.

'Was it worth it?' Calculating costs and benefits

Resources aren't infinite, especially in the current economic context. Policy decisions adjudicate the judicious use of resources and, therefore, will require evidence about 'value for money'. Hence, evaluations often include an element of cost-benefit analysis (CBA).

This is, broadly speaking, a matter of calculating two types of costs – the cost of the problem which the operation is aiming to prevent or reduce, and the cost of the resources which

are required to design and implement the operation – in order to come up with the amount of money saved by the operation (or not).

In practice, this is a complicated exercise, which requires good data, accounting expertise, and the imagination to develop cost measures for what are often intangible items, such as psychological costs and benefits, diplomatic or reputational gains, and so on. A sophisticated CBA will use measures of intensity of inputs and outputs to estimate thresholds at which returns diminish[19].

Choosing an evaluation design: Robustness vs. flexibility

How does one choose an evaluation design among the diversity of options available to evaluators? The answer might seem obvious enough: whichever design can answer the questions and deal with the thorny issues outlined so far in this section.

We want a design that:

- *measures operational impact accurately;*
- *captures the inputs and outputs of the activity faithfully;*
- *eliminates the greatest number of impediments to the establishment of causal attribution (the threats to internal validity;* see Table 2*), as confidently as possible;*
- *provides an understanding of the processes involved in producing the outcome, including the circumstances in which these mechanisms are likely to work again (by ruling out threats to external validity;* see Table 3*).*

Of course, this is easier said than done. If one kind of design were known to achieve all this, there would be no need for this report.

Since such a panacea doesn't exist, the possibilities are as follows. Each has its own pros and cons.

Randomised control trials

Pros. RCTs are the best evaluation designs when it comes to ruling out rival explanations with a high degree of confidence (as long, of course, as they are well-administered).

Randomised experiments, like their names indicate, allocate treatment (the measure delivered by the operation) on a random basis to some targets or sites, but not others. The ones not treated serve as a control group.

[19] See Farrell et al. (2005) for a technical discussion of CBA.

In both groups, quantitative measurements are taken before and after the implementation of the treatment to evaluate effects. Random allocation ensures that members of the treated and control groups do not differ in any other way than whether or not they have received the treatment. Comparing treated and untreated units allows the evaluator to measure the effects of the treatment, and only the effects of the treatment.

If the effects are shown to be significant (after statistical analysis), then the conclusion is that the operation under evaluation, and only the operation, can be responsible for the differences observed between treatment and control groups. RCTs score high on internal validity. In other words, *they are best fit to tackle the challenge of attribution.*

Cons. Before one can carry out an RCT, certain conditions have to be met. Chief among these conditions is that of *independence between the members of the population participating in the trial*. In short, there should be no link or tie between the members that could be responsible for spreading the effect of the treatment. Treating individual A should have no effect on individual B. This is important, since, for obvious reasons the randomly assigned control and treatment groups must remain independent of each other - or the ability to attribute treatment effects with confidence goes up in smoke.

Now let's consider OIAs. In many cases, the IO seeks to exploit precisely the existence of ties between individuals. For example, it wants to diffuse a message among the members of a social network, counting on the fact that the members targeted initially will 'contaminate' the rest.

Let's imagine that one wants to implement an RCT to evaluate the effectiveness of embedding agents of influence in online forums where radicalising activity is known to take place. The agents are tasked with carrying out scripted counter-radicalising activities (e.g. 'friending' forum members; subtly providing alternative viewpoints)[20]. The forums are identified and some are allocated counter-radicalising agents, while others are left alone. Activity is monitored in both groups of forums using the exact same indicators or metrics, for the same period of time.

At the end of the evaluation exercise, could one conclude with a high degree of confidence that a noticeable difference in the amount of radicalising activity between the groups was attributable to the operation? Critics could make a reasonable case that, given the very nature of online radical networks, it is not possible to rule out that the members of one group interacted with the others, unbeknown to the evaluators. Given the nature of online identities, it may even be that some of the same individuals were present in both groups, under different pseudonyms!

On top of the condition of independence, which is incredibly hard to achieve in permeable, online social systems, *there must also be a sufficient number of similar cases to assign randomly to both treatment and control groups*.

In the above example, that means enough forums need to be identified, with similar structures, composition and environment, and the same sort of radicalising activity taking place in all of them. This is a tall order. In some cases, the circumstances of an intervention are so unique or specific that randomisation is simply impossible to consider.

[20] The author thanks Manuel Eisner for providing this hypothetical sketch of an RCT.

As a rule, RCTs thrive on homogeneity. They do better when the members of the population targeted are roughly similar, rather than when each of them is characterised by a set of unique or special circumstances. This preference for uniformity increases statistical confidence (it supports confident attribution; the answer to 'did we do it?'), but it undermines the ability to generalise (the answer to 'what did we learn?).

As Nick Tilley (2009:168) puts it:

"The populations from which cases are randomly assigned [by RCTs] are always, and inevitably, spatio-temporally specific. It cannot logically be concluded that just because an effect is produced among one group at one place and time, it will be experienced in another group at another place and time. This may not matter, in practice, where groups can be assumed to be invariant in relevant respects. But it does matter if this assumption cannot plausibly be made."

Tilley concludes,

"In relation to offenders, victims and offending the assumption of invariance is, at best, highly contestable!"

One can make the case that the same objection will apply to targets of IAs. Short of a *knowledge-based argument* that two populations are similar *in all the ways that matter*, the conclusions of one RCT cannot be transferred over to a new case with confidence.

As the proponents of RCT themselves recognise, the only way to establish whether the effect of an operation is generalisable or replicable in different conditions is to carry out a lot of RCTs in different settings, then subject the findings to systematic review and meta-analysis.

However, to achieve a high level of confidence over the verdict of the evaluation, *RCTs require a high level of control over the environment in which the evaluation activity (and therefore the operation) takes place*. Randomisation, homogeneity…these are artificial conditions, which may be hard to set up in practice. Therefore, carrying out enough RCTs to make a systematic review worthwhile may demand a significant amount of resources and time.

To sum up: RCTs are high-precision tools. Used proficiently, they are highly reliable and can provide robust estimations of the net effects of an operation. The trade-off is that they are intrusive and require specialised training to design and administer. They also work best under conditions which influence activities might be unable to meet. RCTs work best in small, closed systems, with single-measure operations aimed at a well-defined population. While RCTs tackle the challenge of attribution effectively, they do less well against the problem of generalisation. Accounting for the intrinsically open-system nature of OIAs is a tough experimental challenge to meet.

Quasi-experiments

Pros. *Quasi-experiments (QEs) are designs which try to emulate an experimental approach in situations where RCTs cannot be administered.*

These types of designs are often employed in the evaluation of place-based interventions (for example, evaluating the effectiveness of installing alley-gates to prevent residential burglary in a neighbourhood). The logic of using a control group to compare against the treatment group and estimate the size of the intervention's effects remains the same. Sometimes the 'control' is the treated population itself – one simply turns the intervention on, then off, then on again and compares effects over time (this usually requires that the operation go on over a long period).

The closer the control and treatment groups can be matched, and the greater the number of measurements taken over a period of time before and after implementation of the operation, the higher the internal validity of the results – in other words, the stronger the confidence in the design's ability to attribute causal responsibility to the operation for any change.

As a rule, QEs are *less intrusive than RCTs*; they impose fewer artificial conditions and require fewer assumptions, hereby avoiding some of the hurdles to generalisation set up by more stringent experimental designs.

Cons. Quasi-experiments cannot make the same claims to internal validity as RCTs (they do not meet the challenge of attribution as well), though well-crafted QEs can come very close. Because cases are not assigned randomly, they cannot rule out that *some other difference* between treatment and control group is responsible for changes observed, other than the intervention under evaluation. And because they *assume* that they have succeeded in selecting a well-matched, but independent set of control cases, they face the same difficulty as RCTs when operating in open social systems – how to ensure that the control group is not affected by the intervention applied to the treatment group, while still similar enough to the treatment group to be of use?

Neighbourhood-based designs encounter this issue when trying to rule out displacement effects (i.e. the possibility that the problem has been 'pushed over' to another area as a result of intervention in the treated neighbourhood). For this reason, immediately-adjacent neighbourhoods aren't picked as controls. But the further away the control area is located, the more likely it will differ in some (possibly significant) way from the treatment zone, hereby compromising the integrity of the comparative design. Once again, this issue is likely to be of special relevance to the evaluation of OIAs.

As will be obvious from this brief description, QEs, like RCTs, require experience and expertise to design and administer effectively. For this reason, the quality of QEs can vary widely, from sophisticated, multiple-group interrupted time-series designs to simple, small-area before/after set-ups, which struggle to rule out rival explanations.

The ability to meet the attribution challenge varies according to the level of sophistication of the QE.

To sum up: Well-designed QEs can come close to RCTs as far as tackling the challenge of attribution, but in turn they face much of the same problems regarding their ability to deal with

open systems (i.e. intervention and comparison groups must be assumed to be completely independent from one-another). Designing robust QEs can be, to some extent, even more of a challenge than designing robust RCTs. They are, however, more flexible and less intrusive than randomised experiments.

Non-experimental designs

Pros. The pros and cons of non-experimental designs are more difficult to synthesise, as this category houses wildly differing approaches, including simple after-only designs (which only take measures after the intervention), retrospective before/after designs, cross-sectional studies (for example, comparing survey data across different countries using statistical analysis), longitudinal studies, and qualitative designs.

The appeal of non-experimental designs is that they can adapt to the environment in which the operation is conducted, as well as to the resources available to the evaluators (including levels of staffing and training).

Cons. These designs struggle to rule out rival explanations, because they cannot control for the many factors which could be responsible for the changes observed. They are also unable to deliver precise estimates of the effect of the operation – the sort of estimates one might need to calculate costs and benefits.

Nevertheless, realist evaluators, who claim no allegiance to a particular design, will argue that non-experimental designs can serve a valuable purpose, as long as they are guided by hypotheses about the context-mechanism-outcome (C-M-O) configurations at work in the intervention.

They would not, for example, assume the need for randomisation; indeed they might say that purposeful sampling is called for to test hypotheses about which competing mechanisms are likely to be responsible for change at particular sites or among particular kinds of people. They would also see a role for qualitative methods (e.g. focus groups) in evaluation design.

To sum up: Non-experimental designs come in many shapes and vary wildly in their robustness. They can never claim to deal with the problem of attribution effectively. However, given their versatility and low level of intrusiveness, they may have their place in a *strongly theory-guided evaluation*, as a means of investigating specific hypotheses about the effect of particular mechanisms in particular contexts.

Process evaluation designs

Process evaluation designs are add-ons to other types of designs. They document how the operation was implemented and what did or did not go according to plan. (*A note:* things that don't go according to plan are not necessarily bad things; adjustments 'on the fly' may be responsible for a positive outcome – all the more reason to keep track of them.)

Process evaluations keep track of inputs, activities, procedures and timelines. They can even record the state of mind of the implementers – their perspective on what went wrong (or

right). They do not, however, include measures of outcomes, and therefore are not used to measure effects. Their main purpose is to establish accountability.

Table 4 Interpreting results of impact and process evaluations

		Process Evaluation Results	
		Operation implemented as planned, or nearly so	Operation not implemented as planned, or in a radically different manner from what was planned
Impact Evaluation Results	Objective achieved	This is ground to think the operation as planned was a success	This suggests that other factors may be responsible for success, that the operation was 'accidentally' successful, or that success was due to adaption of the operation on the ground
	Objective not achieved	This is ground to think that the operation was ineffective, and that some other kind of operation should be tried	There is little to learn here. If the operation has been implemented as planned, it might have been successful, but it's not possible to say from this evaluation

Source: Adapted from Eck (2005)

They can also strengthen attribution and diagnose the causes of failure. If the plan was followed to a T, a stronger case can be made that the operation was responsible for the outcome; conversely, if the plan wasn't followed, then there's no call to blame the idea behind the operation for its failure (see Table 4). Process evaluations will also play a crucial role in establishing the mechanism(s) at work, allowing the evaluators to draw generalizable lessons from the evaluation.

There are no obvious trade-offs involved in conducting process evaluations, with the exception that they add a layer of administrative accountability to an operation.

To sum up: In combination with an outcome measurement design, process evaluations provide support for conclusions regarding both attribution and generalisation.

In the next section, the argument is put forward that designing evaluations to deal with attribution and generalisation is necessary, but not sufficient to support the development of effective IAs. Evaluations and operations are technological, not scientific, endeavours, and effective technologies must above all be sensitive to user needs.

The choice of evaluation design can only be the result of an exercise which balances optimally the demands of attribution, generalisation and usability.

Further reading

Eck, J.E. (2002). *Assessing Responses to Problems: An Introductory Guide for Police Problem-Solvers* (Updated 2011). Problem-Solving Tool Series. Community Oriented Policing Services, U.S. Department of Justice. Available from: http://www.popcenter.org/tools/assessing_responses/

Tilley, N. (2009). *Crime Prevention*. Cullompton: Willan.

Clarke, R.V. and J.E. Eck (2005). *Crime Analysis for Problem Solvers. In 60 Small Steps*. Washington, D.C.: Office of Community Oriented Policing.

Sampson, R. (2010). "Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology." *Journal of Quantitative Criminology*, 26: 489-500.

Welsh, B.C. and D.P. Farrington (eds.) (2006). *Preventing Crime: What Works for Children, Offenders, Victims, and Places*. New York, NY: Springer.

SECTION 9

Recasting evaluation as a technological endeavour: Overcoming the problem of usability

"The real breakthrough came when tested technological rules could be grounded on scientific knowledge (Bunge, 1967), including law-like relationships from the natural sciences.

For instance, one can design an aeroplane wing on the basis of tested, technological (black box) rules, but such wings can be designed much more efficiently on the basis of tested and grounded technological rules, grounded on the laws and insights of aerodynamics and mechanics."

Van Aken (2004), on the reasons behind the success of the engineering disciplines

Evaluation design, like operation design, is a technological endeavour

As should now be clear, a thread of tension runs through CP between evidence-based and realist philosophies, between tackling the problem of attribution and the challenge of generalisation; between reaching for scientific validity and maximising policy relevance.

This schism isn't unique to CP. The same tug-of-war takes place in other disciplines concerned with translating scientific knowledge into action, such as public health or management science. The friction is more pronounced in areas where the scientific knowledge-base is still maturing, and where no unifying theories or models of explanation dominate, as is the case in the influence domain. While scientific knowledge production in the physical and natural sciences is cumulative, elsewhere it is competitive, new theories challenging older frameworks and new

problems being studied from the ground up.

Does this mean, then, that the evaluation of IAs is condemned to struggle with the self-same tension between basic and practical knowledge, scientific robustness and usability?

The answer is yes.

This is not, however, an insurmountable problem.

Indeed, the case can be made that if the conflict seems intractable, it is only because the problem has been stated improperly.

Despite what the discourse of evaluators in CP and cognate fields might suggest, the role of evaluation is *not* to establish the validity of a scientific truth. That is a job for the scientific method.

Indeed, evaluation *cannot* be a test of a scientific theory or hypothesis, because the object of the evaluation – the operation – 'incarnates' much more than scientific constructs. It also embodies all manner of assumptions regarding the nature and the causes of the problem to be tackled, the specifications of the desired outcome, the principles to follow in order to design solutions, and the tools to use in order to implement them.

IOs are not conducted for the purpose of scientific research, but for their own ends. Ergo, what evaluations assess are the effectiveness of a course of action and of the means chosen to carry it out. It follows that *evaluation is a technological matter,* rather than a scientific one.

Dixit Mario Bunge (2001):

"Technology is the sector of human knowledge concerned with the design and redesign, maintenance, and repair of artificial systems and processes." Hence, "[f]acing a practical problem, taking responsibility for it, and reflecting on the best means to solve it under the known constraints and in the light of the available knowledge and resources, may be regarded as a technological problem."

Here, then, is an explanation for the troublesome tension between different schools of thought in CP and other domains of social action: evaluation is framed as a scientific enterprise[21] – to be shaped by the philosophy and logic of science, the concerns and standards of science, and the methods and tools of science – when in fact it is a technological undertaking.

Why does this distinction matter?

Because technologies, even though they are best built upon the knowledge-base produced by the basic and applied sciences, are not the outcome of a scientific process. They are the

[21] See, for example, Pawson, R. (2003). "Assessing the quality of evidence in evidence-based policy: why, how and when?" ESRC Research Methods Programme. Working Paper N°1. Available from: http://ccsr.ac.uk/methods/projects/buxton/Pawson.pdf.

product of *engineering*[22]. And engineering has its own logic, concerns and tools, shaped to suit its particular ends.

There is a fundamental difference between testing a new treatment and evaluating its implementation in a clinical setting; between testing a new drug and assessing the impact of its release into the population – the ways it might end up being prescribed by medical professionals, used or misused, and the reasons why.

Likewise, *there is a fundamental difference between testing a psychosocial theory of influence* and *evaluating an operation built upon its principles and implemented under 'battle conditions'*.

Building confidence in OIAs requires an R&D programme

Recasting the issue in this way has very concrete implications for the development of robust evaluation frameworks, usable evaluation designs, and efficient IOs.

It means, first and foremost, that effective evaluation technologies – and, inextricably, influence technologies – will be the product of a research and development (R&D) process. Indeed, the R&D process leading to evaluation technologies is, logically, a sub-process in a larger R&D programme leading to efficient influence technologies.

Evaluation is merely one of the stages of the problem-solving cycle, familiar to clinicians everywhere:

1. Identifying the problem and its boundaries
2. Analysing the problem to uncover causal factors and causal processes
3. Formulating a solution
4. Designing an intervention
5. Implementing the intervention
6. Evaluating the outcome
7. Making recommendations

For the purpose of this discussion, the R&D process[23] can be broken down into the following steps:

Statement and analysis of a practical problem →

[22] Mistaking engineering for science is a common category error. For example, we speak of 'rocket scientists', when we really should speak of 'rocket engineers'. There is no such thing as a 'rocket scientist'. See Petroski (2010).

[23] Adapted from Mario Bunge's analysis of the 'technological method'.

Identification of a problem-relevant knowledge-domain →

Synthesis of relevant basic and applied scientific laws or mechanisms →

Synthesis of relevant design principles, rules and methods →

Invention of technological rules grounded in the knowledge synthesised →

Outline of the object or process (artefact), which will act upon the problem →

Detailed blueprint →

Testing (alpha and beta) →

Blueprint revision →

Dissemination of grounded and field-tested technological rules

As Mario Bunge (2001) remarks, this process *"is similar to the scientific method, except that technological tests are tests for efficiency rather than truth."*

This is precisely the heart of the matter.

When we evaluate an operation, we are not trying to establish whether a scientific theory is correct (i.e. 'true'), but *whether a chosen course of action* (which can be represented by the configuration of its goals, means, outcomes and side-effects) *was the right one*.

The last step in the R&D process, the *dissemination of grounded and field-tested technological rules*, is there to ensure that those who have to decide on future courses of actions do not have to start from scratch, in the same way that clinicians do not start from scratch with each new diagnosis, but draw from diagnostic and treatment rules accumulated over time, through the evaluation of clinical practice.

Translating S-M-O patterns into technological rules

The technological rules which result from R&D are general. In other words, they are applicable to *a family of problems*, not just to a specific incident or event (e.g. a 'patient'), which is what makes the process worthwhile.

As defined by Mario Bunge (1967:132), a technological rule is

"an instruction to perform a finite number of acts in a given order and with a given aim."

As Joan van Aken notes, there are two kinds of technological rules. *Algorithmic rules* deliver a predictable result following the completion of a specified number of steps. Pharmacological rules which specify how much of a drug to administer relative to the patient's weight are algorithmic. By contrast, *heuristic rules* have to be translated to be made relevant to the practical problem at hand. So while algorithmic rules are of the form 'To achieve X in

situation Z, do Y," heuristic rules are of the form "To achieve X in situation Z, do *something like Y*."

Needless to say, the kind of rules likely to be produced by a R&D programme in the influence domain will be heuristic in nature. There are no recipes for influence.

As van Aken further observes (2004):

"The indeterminate nature of a heuristic technological rule makes it impossible to prove its effects conclusively, but it can be tested in context, which in turn can lead to sufficient supporting evidence" (emphasis added).

Hence, technological rules are best field-tested through multiple case-study designs, where the cases belong to the same problem family. The effectiveness of the course of action (the operation) *is itself a test of the technological rule upon which it was designed.*

But field-testing is only one part (though an essential one) of the validation process. On its own, field-testing yields 'black-box' statements, of the kind 'yes, this works' or 'no, it doesn't work,' *sans explanation*. Grounding technological rules in scientific knowledge – the laws and causal relationships uncovered by research in basic and applied sciences – is also essential.

We return to the need to supplement empirical findings with accounts of the plausible mechanisms which underpin the results. To 'technologise' operations and their evaluation is not to deny their scientific grounding, but to recognise that they are not, and cannot, be shaped by the scientific knowledge-base alone.

The technological rules produced by R&D activity will address both the nature of problems and their solutions, and the tools or processes used to put them into action.

1. In the first instance, they will formalise the System-Mechanism-Outcome patterns uncovered by the evaluation into heuristic rules of the type *'To achieve an outcome of the kind O_X in a system of type S_Y (with composition C_Y, Environment E_Y, Structure S_Y, and/or Mechanism M_Y), employ a measure of type m_Z to activate a mechanism of the kind M_Z.* An added rule might specify, *'To prevent side-effects of the kind SE_X, measures of the type m_Z should have characteristic c_Z'.*

2. In the second instance, the rules will state which tools or means to employ to assist in the design and implementation of the solution. For example, *'To measure outcomes of type O use scale s', 'To design measures of type m use template t'* or *'To evaluate operations of class Op_W use evaluation design of kind D_W'.*

3. In the third instance, evaluation may, in the long run, generate *meta-technological rules* about the effectiveness of certain classes of solutions against certain families of problems or in certain families of systems. Meta rules may even be uncovered which state, *'Strategic objectives of the kind So_X cannot be achieved through influence operations'*.

Towards the 'routinisation' of influence technologies: Lessons from requirement engineering

The ultimate purpose of R&D activity is to achieve a state where technology can be 'routinised'; where it can be operated by any trained professional in the field.

To use the automobile industry as an analogy, the goal is to get to a situation where anyone with an appropriate licence can operate a car, not just test pilots and Formula 1 drivers. To get to this point, blueprints and prototypes need to be developed, tested (first in laboratory, then in field conditions), and then revised. As described in the R&D schema outlined previously, problems have to be analysed and clearly stated, and relevant knowledge-domains exploited – the sort of activity carried out in this report.

To develop a blueprint, however, one needs more than domain knowledge. Prior to developing and designing any solution, engineers elicit *requirements*.

As stated in systems engineering, requirements can be defined as *the functions that a measure or system must perform, and the range of constraints it must satisfy, in order for the objective to be achieved to the benefit and satisfaction of the user(s) or problem-owner(s).*

Well-elicited requirements allow for the optimisation of a technological solution tailored to the user's specific operational environment, constraints and objectives.

Functional requirements specify what the system should achieve for the user, and what features it should have. In the case of evaluation technologies, a functional system or design should provide the best possible answer to both the questions of attribution and generalisation, given a family of IAs.

Non-functional requirements detail the constraints under which the system is expected to operate, and the target values that various functions are expected to meet. In the case of OIAs, a functional evaluation framework might be expected to deal with imperfect sources of data, limited timescales, a given level of training in operators, a user-defined level of confidence in attribution and generalisation findings, doctrinal constraints, cultural mores, and so on.

Non-functional requirements must be taken into account to ensure that the system, once put in place, doesn't fail due to environmental factors. One example here of non-functional requirement failure would be an evaluation design involving multiple randomised controlled trials, which fails to be implemented because:

- conditions on the ground preclude the required level of control over trial conditions;
- the evaluation activity encounters resistance among users due to its intrusiveness, complexity, and lack of adaptability to the terrain;
- the design is unsuited to the nature of the operations being evaluated;
- it requires specially-educated staff;
- it takes too long to deliver results.

Carrying out R&D activity through systemist formative evaluations

In an R&D context, a formative approach to evaluation can serve as a productive and flexible framework and foundation.

Undertaking *systemist formative evaluations* can help to:

- elaborate well-posed problem statements (i.e. solvable problems, which are appropriately matched to tactical or strategic objectives);
- pinpoint and synthesise the relevant knowledge-base;
- elicit functional and non-functional requirements;
- set out specifications for influence technologies;
- generate explicit S-M-O configurations;
- delineate the pool of realistic interventions; and
- produce technological rules for the design of future evaluation and implementation tools.

In other words, a well thought-out package of systemist formative evaluations of IAs could provide the foundations of a comprehensive and rational R&D programme for influence technologies.

Once the programme has delivered grounded and field-tested technological rules, and confidence in IAs (including OIAs) has been built, a framework for routinised summative evaluations can be produced, to be integrated into the normal cycle of operation design for use by professionals in the field, in the same way that clinical innovations eventually diffuse to practitioners in clinical settings. Should the need for a new family of IOs arise in future, the R&D process can be undertaken again.

Only through such a systematic approach can IO capability-building take place on an organisational level and deliver substantive, game-changing innovation in influence technologies.

The foundations of a systemist formative evaluation process model are laid out in the next section of this report. It builds upon the strengths of the evidence-based and realist evaluation models, while adding the analytical rigour of the systemist approach and the practical methods of requirement and design engineering. It is not intended as a definitive 'recipe'. It will, itself, need to be put to the test of usability. Nevertheless, it offers a next step in the evolution of what Gal McKay and colleagues call a *"substantial concept"* for IAs, which has been missing to date.

Further reading

van Aken, J.E. (2004). "Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules." *Journal of Management Studies*, 41(2):219-246.

Bunge, M. (1967). *Scientific Research. Strategy and Philosophy*. Berlin, New York: Springer-Verlag. Reprinted in Bunge, M. (1998). *Philosophy of Science*. 2 Vols. New Brunswick, NJ: Transaction.

Bunge, M. (2001). "The technologies in philosophy." In M. Mahner (ed.), *Scientific realism: Selected essays of Mario Bunge*. Amherst, NY: Prometheus.

Van Lamsweerde, A. (2009). *Requirement engineering*. Chichester: Wiley.

SECTION 10

A systemist evaluation blueprint for online influence activities

"Influence has become the 'must have' accessory for the battlefield. Good. But think how difficult it is to influence, say, your teenage kids, into a particular course of action. You know them. They have grown up in your house. You know the groups they belong to, their interests, their likes and dislikes. Yet as every parent knows influencing a 16 year old into a particular course of action can be difficult.

Now apply this thinking to an Afghan whom you do not know, who has grown up in a completely different culture with different values and beliefs anchored in a wholly different world from our own. You want to influence them? Wow! This is hard stuff to do and whilst the UK's capability and understanding has leapt forward in the last couple of years there is still much work to do."

Rowland and Tatham, 2008

Designing systemist formative evaluations for OIAs

This section sets out the foundations of a systemist evaluation process model, to be implemented as part of a formative evaluation programme for IOs.

As stated, formative evaluations are conducted at the *developmental stage* of a new kind of activity, when there is insufficient knowledge about what sorts of operations might be effective, or even what, precisely, the activity is setting out to achieve (the problem statement).

Formative evaluations encourage structured self-reflection about the nature of problems and ultimate objectives, as well as the development of innovative solutions. The process is very similar to general-purpose requirements elicitation processes in systems engineering.

Formative evaluations are intensive exercises. They can be conducted 'table-top', though they are most effective when carried out alongside live operations. Evaluators, designers and implementers must work closely together to produce a

clear picture of problems and objectives, elaborate appropriate responses, resolve conflicting requirements issues, anticipate implementation hurdles, set up monitoring procedures, and analyse results, in order to disseminate lessons learned in the form of technological rules.

The task of designing and agreeing the formative evaluation process is itself a collaborative undertaking. Evaluators and users must agree what, when and how they are to elicit information out of each other. Given the nature of the formative process, it is often iterative, and may require that evaluators and users go back and forth between phases of elicitation.

Hence, what follows can only be taken as a first draft.

Since the main aim of the proposed formative evaluation process is to bring to light specific kinds of R&D activity needed to deliver and support mature influence technologies, plausible R&D activities are suggested for each phase. However, the list is by no means exhaustive.

There is, indeed, *"still much work to do."*

Table 5 Blueprint of a systemist formative evaluation framework for IAs

	Evaluation Phase	Description	Associated R&D Activities
	Domain scanning	Acquisition of general knowledge of the area and circumstances in which the IO is to be rolled out.	
ELICITATION OF REQUIREMENTS	**Problem statement**	Elaboration of a problem statement, which : 1) is well-posed (i.e. it confirms that the problem exists; it establishes that the issue identified by the end-user is indeed the problem, instead of, for example, a symptom; it defines the problem specifically enough that there is confidence that everyone shares the same understanding); 2) ascertains that the problem identified is solvable in a measurable way; 3) provides a clear picture of the *who, what, where, when, how* of the problem; 4) specifies clearly how the problem or issue is related to overarching goals (e.g. strategic aims), and its relative importance 5) establishes clear aims and objectives for the IO; 6) is agreed with the problem-owner and end-users.	• Design and validate tools and techniques to guide problem statement elicitation and visual representation.
	Problem analysis	Identification and collection information on the stated problem, and analysis data, in order to: 1) narrow the scope of the problem; 2) assess existing responses to the problem, if any have already	• Design and validate tools and techniques to guide and structure problem analysis. • Conduct systemic literature reviews and syntheses of

		been implemented; 3) acquire and synthesise the relevant knowledge-base; 4) identify the causes and markers associated with the problem at all relevant levels of analysis; 5) formulate working hypotheses as to what the key causes of the problem are and what factors or processes might be manipulated to effect change.	behaviour-change research in cognate and commensurate problem domains, in order to expand the knowledge-base beyond attitude-based models of influence. For an example of a systemic synthesis, see Bouhana and Wikström (2011), where a multilevel, systemic model of the causes of radicalisation is produced from the synthesis of disparate studies of Al-Qaeda-influenced radicalisation.
	Target systemic analysis	Systemic analysis of the IO's target (eg. individual, group, community, society), with attention to each of the four systemic components: C (composition); E (environment); S (structure), and; M (mechanisms), in order to uncover, notably: 1) key systems in which the target is embedded, and their characteristics; 2) influence pathways (bonds and mechanisms) between the target and the systems to which it belongs, which can be exploited by the IO; 3) existing sources of influence, which may or may not have a competing agenda and should be either exploited or countered by the IO.	• Design and validate tools to guide and structure target systemic analysis.
	S-M-O hypotheses	Generation of System-Mechanism-Outcome configuration hypotheses, based on the findings of the problem and target systemic analyses.	• Design and validate tools to guide and structure the formulation of S-M-O hypotheses. • Synthesise literature in cognate and

			commensurate problem domains to extract and catalogue documented or theorised S-M-O configurations. This activity can be carried out as part of the systemic syntheses described above.
	Measure design	Design of the specific measures or activities (ie. interventions) which the IO will implement, given the S-M-O hypotheses. Measures should be designed to activate one or more mechanisms, given the systemic features highlighted by the target systemic analysis and the knowledge-base synthesis carried out during the problem analysis phase. The design process should make reference to the contextual features under which the measure is expected to activate the hypothesised mechanism and produce the desired outcome. This includes reference to timing (when the activity should be implemented and for how long). It is also likely to include reference to the agent who will enact the activity (as the source of influence is likely to be a 'key ingredient').	• Design and validate tools to guide and structure the design of activities.
	IO context analysis	Elicitation of information about the IO and its environment, with the aim to establish: 1) the personnel who will implementing the IO (their number, location, training, and so on); 2) the actors whose collaboration may be required to implement the operation ('green' forces, businesses, NGOs, civilian groups, and so on); 3) the IO's timeline (eg. how long can the IO run; when are results expected; is it consistent with what the knowledge-base is the time needed for influence to take effect);	• Design and validate tools to guide and structure the operation context analysis and the elicitation of operation specifications.

		4) the IO's environment (including chain of command) and any other implementation constraints the IO will have to contend with (including constraints on evaluation activity); 5) the specifications which must be met before the IO gets underway (eg. human, material and technical resources), given all of the above; 6) whether any of the activities or measures previously designed must be adapted or refined given operational constraints.	
	IO blueprint	Design of the IO, based on the synthesis of requirements elicited up to this point, taking into account specifications relative to the measures to be implemented and to the context in which the operation will be carried out.	• Design and validate tools to guide and structure the design of the operation blueprint, such as operation matrices or templates.
	Evaluation design	Design of the evaluation, taking into account the IO's blueprint and all elicited requirements and corresponding specifications. The design should contain both *process* and *outcome* evaluation components. Given implementation constraints and user needs, the choice of evaluation design (experimental, quasi-experimental, non-experimental) and evaluation metrics should balance optimally: 1) attribution; 2) generalisation; 3) analysis; 4) usability.	• Design and validate tools to guide and structure the design of evaluations, such as design guides, checklists, templates, and catalogues of best-practice examples in cognate and commensurate domains. The purpose of these tools must be to facilitate the collaboration of evaluators and users, in order to produce an evaluation design which handles trade-offs between the four challenges of evaluation in an optimal way.
	Testing - Alpha	Assessment of the operation blueprint and evaluation design using table-top methodology. This is akin to 'red teaming'. The purpose is to troubleshoot and refine the IO and evaluation design before implementation in the	• Design and validate alpha-testing methodology.

	field, as well as anticipate unintended outcomes of the activity.	
Testing - Beta	Implementation and evaluation the IO against one (or a series of) test field-cases.	• Design evaluation data collection and storage system.
Process and outcome evaluation synthesis	Aggregation and interpretation of the evaluation findings into the following products: 1) synthesis of the findings of the process and outcome evaluations (ie. interpretation of the IO's impact in terms of context; eventual diagnosis of implementation of theory failure); 2) validation or invalidation of S-M-O hypotheses; 3) systemist synthesis of findings in the event of multiple case studies; 4) derivation of general technological rules from the synthesis.	• Design and validate methodologies and tools for evaluation synthesis and the formulation of standardised evaluation products.
Dissemination of technological findings	Dissemination of the conclusions of the formative evaluation in the form of: 1) heuristic rules of influence; 2) heuristic rules of implementation; 3) analytical tools; 4) standardised evaluation designs; 5) training materials and other documentation of lessons learned and best practice.	• Design and validate dissemination processes, including training.

SECTION 11

Conclusion: What future for influence activities?

The author was tasked with outlining the foundations of a framework for the evaluation of OIAs – a first step towards building greater confidence in the effectiveness and value of IAs conducted in a military context.

The task was taken a little further. This report presents the barebones of an R&D programme, in reconnaissance that *all four challenges of evaluation – attribution, generalisation, analysis, and usability – must be tackled* if influence technologies are to become more efficient and more reliable.

Nevertheless, this is not the only possible way forward. It would be perfectly feasible to implement a more modest approach to evaluation, of the kind outlined in the NATO report cited earlier on, concerned with the design of functional MoEs rather than with the evaluation of the systemic logic of IOs.

Without going to such a theory-free extreme, one could adopt the pragmatic, 'realism-lite' approach embodied by the Problem-Solving Tools series published by the US Centre for Problem-Oriented Policing. In "Assessing Responses to Problems," criminologist John Eck provides CP practitioners with a basic introduction to the logic of realist evaluation, as well as techniques

"Even during wars of national survival or the destruction of WMD, conflict will remain focused on influencing people. The battle of the narratives will be key, and the UK must conduct protracted influence activity, coordinated centrally and executed locally."

Development, Concepts and Doctrine Centre, 2010

"Progress in human affairs, whether in science or in history or in society, has come mainly through the bold readiness of human beings not to confine themselves to seeking piecemeal improvements in the way things are done, but to present fundamental challenges in the name of reason to the current way of doing things and to the avowed or hidden assumptions on which it rests."

E.H. Carr, 1961

of response-assessment, including some standard, minimally intrusive evaluation designs.

It would not take too much work to adapt such a guide to the needs of practitioners in the OIA domain – though suitable MoEs would have to be trialled. To say that nothing of value would be gained from implementing even a basic approach to evaluation would be disingenuous, if only because there is an inherent benefit to the introduction of practices that encourage self-reflection on the part of operation designers. Too many courses of action are undertaken without a clear idea of what the goals are or what success would look like.

However, this is not the approach to take if one is aiming for capability-building on any significant scale. Whether such capability-building is desirable is, of course, a question for doctrine and policy-making – a discussion to which this author, a criminologist naive in these matters, has little to contribute.

That said, if it is proposed – as one might gather from publications such as the DCDC's *The Future Character of Conflict* – that IAs should play a larger part in the overall context of UK operations, then a more ambitious agenda is called for.

If *"protracted influence activities"* are indeed the endgame, then they must be supported by a substantial and rational R&D programme, in the same way that support for kinetic operations requires large-scale R&D programmes. No one thought twice about the need for R&D in order to adapt to new battlefield tactics (such as the proliferation of improvised explosive devices), to improve battlefield medicine, or to counter new and emerging cyber-threats.

Yet, as Gal McKay and colleagues point out, IOs are still conducted at the same level of technological development which characterised operations ninety years ago.

The R&D logic advocated here is the same logic which has motivated development in other areas of operations. Functioning, reliable, cost-effective influence technologies capable of long-term strategic deployment cannot be developed *ad hoc*.

The key question is *'How important is it to build confidence in IAs?'*, followed by its corollary, *'How do we ensure that this confidence is, and remains, well-placed?'*

This report has drawn from experience in CP to make the case for a systemist approach to evaluation. As stated at the outset, the point was not to claim that CP evaluators have it all figured out; instead, the idea was to capitalise on weaknesses as well as strengths, building on the latter while proposing constructive ways to address the former.

As this reports hopes to have shown, there is ground to stand on between rigorous, but impractical scientific evaluation designs; realist, but analytically fuzzy frameworks; and *ad hoc*, low-maintenance, but unreliable assessment tools.

As per the advice of historian H.E. Carr, this ground was uncovered by questioning the *"hidden assumptions"* upon which rest not only evaluation activities, but social interventions more generally, of which CP and IAs are two instances:

1. The assumption that two of the functions of evaluation activity (establishing attribution and providing grounds for generalisation) should supersede the requirement of usability; and
2. The assumption that evaluation, and to some extent the activities being evaluated, should be treated as scientific undertakings, with all the philosophical and methodological trappings this entails.

As to the first assumption, the argument was put forward that usability is just as important as attribution and generalisation, because what is unusable will not, by definition, be put to use, in which case concerning ourselves with the problems of attribution and generalisation is moot.

As to the second, the case was made that to treat evaluation and influence activities as scientific products is to commit a category error. They are technologies and should be handled as such, meaning that their development and assessment should be part of a full problem-cycle, which includes the elicitation of user requirements.

Taking an engineering approach to influence technologies will, inevitably, lead to confronting other (more or less hidden) assumptions and to the formation of recommendations, some of which are already actionable:

1. It will challenge the idea that operation designers can go straight to theoretical models and empirical research in the basic or applied sciences – such as social psychology, social networking or decision theories – and put these findings to use without further ado.

   To do this is to ignore the fact that, while engineers build upon established scientific principles, they also bring to bear a body of knowledge unique to their own discipline: an understanding of systems and design principles, which is not contained within the scientific corpus itself, but is validated through engineering's own methods and processes.

   Science is about understanding, while engineering is about problem-solving. *Without the application of latter, the former cannot deliver efficient action.*

   It takes more than knowledge of the laws of physics to build a bridge, and it takes more than psychological principles, behavioural models or decision theories to design IOs.

   Recommendation: When commissioning, soliciting or turning to the products of research and theoretical development in the human and social sciences in the context of OIAs, keep in mind that these products need to be assessed against user requirements, both functional and non-functional, as would any other new technological system, prior to implementation.

2. Rigorous problem analysis and the subsequent synthesis of relevant knowledge-bases are likely to challenge the received wisdom that IAs are and should be chiefly about changing 'attitudes'. A shift towards multi-level, integrated behavioural models is likely to take place, to reflect the state-of-the-art in the behavioural sciences. The notions of behavioural change which underpin current thinking on influence have their roots in thirty-year-old literature and need a significant update. New models are needed which articulate developmental causes of propensity acquisition as well as situational causes of behaviour, taking into account social ecological and systemic (contextual) processes.

   Importantly, not all 'theories of influence' or 'behavioural change' are equal or can be of use to inform influence activity. The assumptions upon which influence models are based in some domains, such as in marketing communication, are significantly different from those that underpin influence operations, which should caution users against importing techniques from other field wholesale.

   R&D activity may reveal ultimately that investment in basic science is required before influence models fit to drive OIAs can emerge. It is here that experimental designs are most likely to benefit the knowledge-base.

   Recommendation: Commission systemic syntheses of the literature on behavioural change in commensurate domains, which reflect the state-of-the-art in social environmental and ecological science, social cognitive neuroscience, and other systemic understandings of human behaviour, in order to generate new analytical frameworks for IO design, which do not rely on outdated attitude-change models.

3. Finally, a systemist outlook can only challenge expectations, if any remain, that IOs can achieve their objectives regardless of what goes on 'in the field'. In open social systems, actions are as loud as words, if not louder.

   Recommendation: Building confidence in IAs also means *managing expectations of what they can achieve*.

As far as the approach proposed here, the next stage is to subject the draft of systemist evaluation process model to further development, alpha-testing and fine-tuning. As it is itself a technology, it needs to be adapted to the requirements, mores and values of the user organisation. Once a prototype is ready, it may be piloted upon planned, or, if impractical, historical or simulated IOs, as a whole or in part.

This would constitute first a step towards devising a coherent R&D programme for influence technologies. Whether that course of action is desirable is not for the author to say. The present report can only aim to inform that decision.

References

Alison, L., Bennell, C., Mokros, A. and D. Ormerod (2002). "The personality paradox in offender profiling: A theoretical review of the processes involved in deriving background characteristics from crime scene actions." *Psychology, Public Policy and Law* , 8(1): 115–35.

Astbury, B.and F.L. Leeuw (2010). "Unpacking black boxes: Mechanisms and theory building in evaluation." *American Journal of Evaluation*, 31(3): 363-381.

Bewick, B., Trusler, K., Barkham, M., Hill, A., Cahill, J., and B. Mulhern (2008). "The effectiveness of web-based interventions designed to decrease alcohol consumption: A systematic review." *Preventative Medicine*, 47:17-26.

Blamey, A. and M. Mackenzie (2007). "Theories of change and realistic evaluation: Peas in a pod or apples and oranges?" *Evaluation*, 13: 439-455.

Bouhana, N. and P-O. Wikström (2011). *Al-Qaeda-influenced radicalisation: A rapid evidence assessment guided by Situational Action Theory*. RDS Occasional Paper 97. London: Home Office. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/116724/occ97.pdf.

Bowers, K. and S. Johnson (2005). "Using Publicity for Preventive Purposes." In Nick Tilley (ed.) *Handbook of Crime Prevention: Theory, Policy and Practice*. London: Willan.

Bowers, K.J., and Johnson, S.D., and Hirschfield, A.F.G (2004). The measurement of crime prevention intensity and its impact on levels of crime. *The British Journal of Criminology*, 44(3), 1-22.

Braga, A.A. and D. Weisburd (2006). *Police Innovation: Contrasting Perspectives.* Cambridge: Cambridge University Press.

Bunge, M. (1967). *Scientific Research. Strategy and Philosophy*. Berlin, New York: Springer-Verlag. Reprinted in Bunge, M. (1998). *Philosophy of Science*. 2 Vols. New Brunswick, NJ: Transaction.

Bunge, M. (1996). *Finding Philosophy in Social Science*. Yale, Mass: Yale University Press.

Bunge, M. (2001). "The technologies in philosophy." In M. Mahner (ed.), *Scientific realism: Selected essays of Mario Bunge*. Amherst, NY: Prometheus.

Bunge, M. (2004). "How Does It Work? The Search for Explanatory Mechanisms." *Philosophy of the Social Sciences*, 34(2): 182-210. Available from:

http://www.gemas.fr/dphan/cosmagems/docs/socio/PhilosophyOfTheSocialSciences2004Symposium_2Bunge.pdf

Bunge, M. (2006). "A systemic perspective on crime." In P-O Wikstrom and R. Sampson, *The explanation of crime: Context, mechanisms and development*. Cambridge: Cambridge University Press.

Bunge, Mario (2006). *Chasing reality: Strife over realism.* Toronto: University of Toronto Press.

Cabinet Office (2011). *The UK Cyber Security Strategy: Protecting and promoting the UK in a digital world.* London: Cabinet Office.

Carr, E.H. (1961). *What Is History?* London: Penguin.

Clarke, R.V. and J.E. Eck (2005). Crime Analysis for Problem Solvers. In 60 Small Steps. Washington, D.C.: Office of Community Oriented Policing.

Collings, D. & Rohozinski, R. (2006). *Shifting Fire: Information Effects in Counterinsurgency and Stability Operations.* Carlisle: US Army War College. Available from: http://www.au.af.mil/au/awc/awcgate/army-usawc/shifting_fire.pdf.

Development, Concepts and Doctrine Centre (DCDC) (2010). *The Future character of conflict*. DCDC Strategic Trends Series. Available from: https://www.gov.uk/government/publications/future-character-of-conflict.

Eck, J.E. (2002). Assessing Responses to Problems: An Introductory Guide for Police Problem-Solvers (Updated 2011). Problem-Solving Tool Series. Community Oriented Policing Services, U.S. Department of Justice. Available from: http://www.popcenter.org/tools/assessing_responses/.

Eck, J.E. (2005). "Evaluation for Lesson Learning." In N. Tilley (ed.), *Handbook of Crime Prevention and Community Safety*. London: Willan.

Eck, J.E. (2006). "When is a bologna sandwich better than sex? A defence of small-n case study evaluation." *Journal of Experimental Criminology*, 2:345-362.

Ellefsen, B. (2011). "Evaluating Crime Prevention: Scientific Rationality or Governmentality?" *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 12(2):103-127.

Farrell, G., Bowers, K., & Johnson, S.D. (2005). "Making Cost-Benefit Analysis Useful for Criminal Justice Evaluations by using a Limited Portfolio of Benefit-Cost Ratios: A Framework and an Example using Evidence from the Burglary Reduction Initiative." In M. Smith and N. Tilley (eds.) *Crime Science: New Approaches to Preventing and Detecting Crime*. London: Willan.

Glasman, L.R. and D. Albarracín (2006). "Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation." *Psychological Bulletin*, 132(5): 778-822.

Gottfredson, M. R. and T. Hirschi (1990). *A General Theory of Crime*. Stanford, CA: Stanford University Press.

Hodgson, G.M. (2012). "On the Limits of Rational Choice Theory." *Economic Thought*, 1:94-108.

Hutchinson, W. (2010). "Influence operations: Action and attitude." Proceedings of the 11th Australian Information Warfare and Security Conference, Edith Cowan University, Perth

Western Australia, 30th November - 2nd December 2010. Available from: http://ro.ecu.edu.au/isw/33/.

Kahneman, D. (2011). *Thinking fast and slow*. London: Penguin.

Keller, R.(2010). *Influence Operations and the internet: A 21st Century issue. Legal, doctrinal and policy challenges in the cyber world*. U.S. Air University: Air War College. Available from: http://www.au.af.mil/au/awc/awcgate/maxwell/mp52.pdf.

Larson, E. V. Darilek, R.E., Gibran, D., Nichiporuk, B., Richardson, A., Schwartz, L. H., and Quantic-Thurston, C. (2009). *Foundations of Effective Influence Operations: A Framework for Enhancing Army Capabilities*. RAND Corporation. Available from: http://www.rand.org/pubs/monographs/MG654.html.

Lieberson, S. and J. Horwich (2008). "Implication analysis: A pragmatic proposal for linking theory and data in the social sciences." *Sociological Methodology*, 38(1): 1-50.

Lustria, M. L., Cortese, J., Noar, S., and R. Glueckauf (2009). "Computer- tailored health interventions delivered over the web: Review and analysis of key components." *Patient Education and Counselling* , 74:156-173.

McKay, A., Tatham, S. and L. Rowland (2012). *The Effectiveness of US Military Information Operations in Afghanistan 2001-2010: Why RAND missed the point.* Central Asia Series. UK Defence Academy. Available from: http://www.da.mod.uk/publications/library/central-asian-series/20121214_Whyrandmissedthepoint_U_1202a.pdf

Merton, R.K. (1936). "The unanticipated consequences of purposive social action." *American sociological review*, 1(6): 894-904.

Munoz, A. (2012). *U.S. Military Information Operations in Afghanistan: Effectiveness of Psychological Operations 2001–2010*. Santa Monica, CA: RAND, National Research Defense Institute. Available from: http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1060.pdf.

Murphy, D.M. (2012). "The future of influence in warfare." *Joint Force Quarterly*, 64: 47-51. Available from: http://www.ndu.edu/press/lib/pdf/jfq-64/JFQ-64_47-51_Murphy.pdf.

NATO (2011). *How to Improve your Aim: Measuring the Effectiveness of Activities that Influence Attitudes and Behaviors.* RTO Technical Report. TR-HFM-160. Available from: http://info.publicintelligence.net/NATO-MeasuringInfluence.pdf.

Pahlavi, P. C. (2007). "The 33 Day War: An Example of Psychological Warfare in the Information Age." *Canadian Army Journal*, 10:12-24. Available from: http://www.army.forces.gc.ca/caj/documents/vol_10/iss_2/CAJ_vol10.2_05_e.pdf.

Pawson, R. (2006). *Evidence-Based Policy: A Realist Perspective*. London: Sage.

Pawson, R. and N. Tilley (1997). *Realistic Evaluation*. London: Sage.

Pawson, R. and N. Tilley (2004). *Realist Evaluation*. Available from: http://www.communitymatters.com.au/RE_chapter.pdf

Perry, R.L. (2008). "A multidimensional model for PSYOP measures of effectiveness." *IOSphere*, 9-13. Available from: http://www.au.af.mil/info-ops/iosphere/08spring/iosphere_spring08_perry.pdf.

Petroski, H. (2010). "Engineering is not science – And confusing the two keeps us from solving the problems of the world." Available from: http://spectrum.ieee.org/at-work/tech-careers/engineering-is-not-science.

Rate, C.R. (2011). *Can't count it, can't change it: Assessing influence operations effectiveness*. Carlisle, PA: U.S. Army War College. Available from: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA560244.

Rootman, I., Goodstadt, M., Hyndman, B., McQueen, D., Potvin, L., Springett, J. and E. Ziglio (eds.) (2001). *Evaluation in Health Promotion: Principles and Perspectives*. WHO Regional Publications. European Series No. 92. Available from: http://www.euro.who.int/__data/assets/pdf_file/0007/108934/E73455.pdf.

Rowland, L. and S. Tatham (2008). *Strategic Communication & Influence Operations: Do We Really Get It?* Special Series. UK Defence Academy. Available from: http://kingsofwar.org.uk/wp-content/uploads/2010/09/RowlandsTathamInfluencepaper1.pdf.

Sampson, R. (2010). "Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology." Journal of Quantitative Criminology, 26: 489-500.

Sherman, L. (1998). *Evidence-Based Policing*. Washington, DC: Police Foundation. Available from: http://www.policefoundation.org/content/evidence-based-policing

Thomas, T.L. (2007). "Hezballah, Israel and cyber PSYOP." *IOSphere,* 31-35. Available from: http://fmso.leavenworth.army.mil/documents/new-psyop.pdf.

Tilley, N. (2009). *Crime Prevention*. Cullompton: Willan.

van Aken, J.E. (2004). "Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules." *Journal of Management Studies*, 41(2):219-246.

van Lamsweerde, A. (2009). *Requirement engineering*. Chichester: Wiley.

Welsh, B.C. and D.P. Farrington (eds.) (2006). Preventing Crime: What Works for Children, Offenders, Victims, and Places. New York, NY: Springer.

Wikström, P-O. (2007). "Doing Without Knowing: Common Pitfalls in Crime Prevention." In Farrell, G., Bowers, K., Johnson, S. and M. Townsley (eds.), *Imagination for Crime Prevention: Essays in Honour of Ken Pease*. Crime Prevention Studies Vol. 21.

Wikström, P-O. (2011). *"Does Everything Matter? Addressing the Problem of Causation and Explanation in the Study of Crime."* In McGloin. J.M., Sullivan, C. J. and L.W. Kennedy (eds), *When Crime Appears: The Role of Emergence* London. Routledge.

Welsh, B.C. and D.P. Farrington (2006). *Evidence-Based Crime Prevention*. In B.C. Welsh and D.P. Farrington (eds.), *Preventing Crime: What Works for Children, Offenders, Victims and Places*. New York, NY: Springer.

About the author

Dr. Noémie Bouhana (Author)

Dr. Bouhana is a lecturer at University College London, where she leads the Counter Terrorism Research Group of the Department of Security and Crime Science, Faculty of Engineering Science, and directs the MSc Programme in Countering Organised Crime and Terrorism. Her research is chiefly concerned with uncovering the systemic causal processes involved in the emergence of individual radicalisation.

Email: n.bouhana@ucl.ac.uk

Ms. Kate Gibson (Research Assistant)

Ms. Gibson is a doctoral researcher at University College London, Department of Security and Crime Science. She is carrying out her doctoral research project on internal firearms procurement networks in the United Kingdom. Prior to that, she accumulated six years of experience undertaking research projects for clients including the Home Office, Cabinet Office and the Ministry of Justice.

Department of Security and Crime Science

University College London

35 Tavistock Square

London WC1H 9EZ

+44 (0)20 3108 3206