skema THINK TANK | PUBLIKA

NOTE
UNCERTAINTIES

MAPS WITHOUT DESTINATIONS?

June 2026

From describing information flows to measuring effects in digital influence

Jean LANGLOIS-BERTHELOT & Paul JANIN

SKEMA PUBLIKA

SKEMA Publika is SKEMA Business School's international think tank. It analyses major social, economic, technological and geopolitical changes to inform public and private decision-making. Drawing on the school's research and on scientifically validated external contributions, the think tank fuels public debate and issues recommendations for national and international decision-makers.

SKEMA Publika takes an interdisciplinary and international approach, enriched by SKEMA's global network of campuses and a community of experts from the academic and professional worlds. This international dimension is not simply a network, but a way of thinking. SKEMA Publika therefore connects local dynamics with global transformations to provide a decentralised and multipolar perspective on the major challenges of our time, rather than one centred on a single national lens.

Cover: A very large screen showing many types of television screens. Eamonn Wang. 2023. Unsplash.com

SKEMA Publika
SKEMA Business School, Grand Paris Campus
5 Quai Marcel Dassault – CS 90067
92156 Suresnes Cedex, France

Tel.: +33 (0)1 71 13 39 32
Email: publika@skema.edu
Website: www.publika.skema.edu

UNCERTAINTIES COLLECTION

'Exploring the grey areas of geopolitics, digital technology and global risks.'

The publications in the 'Uncertainties' series analyse tensions, emerging risks and dynamics of instability globally.

EDITORS

Frédérique Vidal is Director of Development at SKEMA Publika and Director of Strategy and Scientific Impact at SKEMA Business School. A full professor of Biology, she was president of the University of Nice Sophia Antipolis from 2012 to 2017, and subsequently served as minister of higher education, research and innovation in the Philippe and Castex governments from 2017 to 2022. She served as special advisor to the President of EFMD and is currently also the Permanent Representative of the Principality of Monaco to the United Nations Environment Programme and the Whaling Commission.

Sean Scull, Think Tank Project Manager, is a doctoral candidate in Information and Communication Science at Université Paul Valéry – Montpellier III. He holds a degree in Political Science with a specialisation in International Relations from the University of Gothenburg, and a Master's degree in International Politics with a focus on Anglophone politics from the University of Toulon. Sean has lived and worked in Sweden and the United States.

AUTHORS

Jean Langlois-Berthelot works at Kannon Labs, where he focuses on simulation, decision intelligence and technology assessment applied to strategic environments. He is currently a visiting professor at the Centre for Advanced Studies at the University of Palermo and is working on experiments conducted as part of the Orion 26 military exercise.

Paul Janin is a senior officer in the French Army, specialising in cognitive science, influence and cognitive warfare. As part of the Army Higher Military Scientific and Technical Education programme, he was tasked with analysing the scientific, academic and institutional ecosystem involved in these fields, carrying out fieldwork, analysis and exchanges with the sector's leading institutions. He notably spent time working at CEA Paris-Saclay, liaising with scientists involved in Defence Innovation Agency (AID) and Directorate General of Armaments (DGA) funding programmes relating to cognitive warfare and influence. He has published in several journals, including Revue Défense Nationale, Polytechnique Insights and ISTE OpenScience.

ABSTRACT

Digital influence has become a central issue in the cyber field. It features in military doctrines on cyber influence operations, in public policies addressing information manipulation, in research on digital platforms, in debates surrounding generative artificial intelligence, and in discussions on democratic resilience. This centrality is no accident: contemporary operations no longer clearly separate technical attacks, information operations and social disruption.

This observation led the authors to formulate the following research question: what are the limitations of information monitoring tools when it comes to measuring the actual effectiveness of an influence operation? Put differently, how can we move from describing information flows to measuring effects? The argument advanced here is straightforward: digital influence can only be assessed rigorously through behavioural change measured against a counterfactual scenario. Narratives, networks, cognitive biases, social intermediaries and platform metrics are useful but intermediate variables. They only become meaningful when they help to estimate what actually changes compared with what would probably have occurred in the absence of intervention.

TABLE OF CONTENTS

INTRODUCTION

Digital influence has become a central issue in the cyber field. It features in military doctrines on cyber influence operations, in public policies addressing information manipulation, in research on digital platforms, in debates surrounding generative artificial intelligence, and in discussions on democratic resilience. This centrality is no accident: contemporary operations no longer clearly separate technical attacks, information operations and social disruption.

A data breach may be exploited to produce reputational effects, a smear campaign may accompany a cyberattack, a symbolic act of sabotage may be amplified through social media, and AI-generated content may accelerate information overload before its accuracy can even be verified. Cyberspace is therefore no longer merely a space of intrusion, compromise or technical defence. It has also become a space for producing cognitive, social and political effects.

This evolution helps explain the proliferation of monitoring, detection and analytical tools. Organisations are now capable of monitoring large volumes of messages, detecting coordination, mapping networks, tracking hashtags, comparing narratives, identifying intermediaries, measuring interactions and producing dashboards. France has formalised this concern in its National Strategy for Countering Foreign Information Manipulation 2026–2030, published on 11 February 2026 by the General Secretariat for Defence and National Security (SGDSN).[1] Yet the increasing deployment of these tools does not resolve the core scientific problem. On the contrary, it may even obscure it.

This brings us to the following research question: what are the limitations of information monitoring tools when it comes to measuring the actual effectiveness of an influence operation? Put differently, how can we move from describing information flows to measuring effects? The indicators currently available (impressions, engagement, repetition of narratives, number of accounts, speed of dissemination and narrative reach) describe an information environment, but they do not directly measure the effect of

[1] Secrétariat général de la défense et de la sécurité nationale (SGDSN), *Stratégie nationale de lutte contre les manipulations de l'information 2026-2030*, 11 February 2026.

an operation. They indicate that content has been seen, shared, commented on or amplified, but they do not reveal whether a population has altered its behaviour, a firmly held intention, a decision, a form of cooperation, a mobilisation effort, abstention, a purchasing decision, institutional trust, or a security practice.

The argument advanced here is as follows: digital influence can only be assessed rigorously through behavioural change measured against a counterfactual scenario. Narratives, networks, cognitive biases, social intermediaries and platform metrics are useful but intermediate variables. They only become meaningful when they help to estimate what actually changes compared with what would probably have occurred in the absence of intervention.

I. MEASURING THE IMPACT OF INFLUENCE OPERATIONS: DISTINGUISHING BETWEEN VISIBILITY AND EFFECT

This issue had already been identified as an operational challenge in a 2018 presentation of findings by the AID and the Human and Artificial Cognition Laboratory at the École Pratique des Hautes Études (EPHE-PSL): influence analysis systems rely largely on variables that are readily available because they are observable, yet are insufficient to establish behavioural impact. There is therefore nothing radically new about the current formulation of the problem. Rather, it confirms a longstanding finding: the field is becoming increasingly adept at describing signals, but still struggles to estimate effects. This distinction is particularly important in the cyber domain, where decisions are made under time pressure, on the basis of incomplete data and with the risk of amplification. Responding too quickly to a hostile narrative may neutralise it, but it could also give it visibility it would never otherwise have achieved. Conversely, remaining silent may be strategically sound, but it can also allow a narrative to take hold. The question therefore must not be: 'Is the narrative spreading?' Instead, it must be: 'Which course of action produces the greatest net behavioural effect compared with inaction and the other options available?'

The scientific literature helps to explain why this is a difficult question. Since the 1940s, research on influence has progressively addressed several dimensions of the problem without ever fully resolving the challenge of measuring causal effects. The first major phase was the study of mass communication. In 1948, Harold Lasswell formulated[2] the now-classic question: who says what, through which channel, to whom, and with what effect? This formulation has the merit of placing effect at the heart of the problem, but it does not yet provide a robust method for estimating it. It frames the question rather than resolving it.

[2] Harold D. Lasswell, 'The Structure and Function of Communication in Society', in Lyman Bryson (dir.), The Communication of Ideas, New York, Harper & Brothers, 1948, pp. 37-51.

The second major phase emerged between 1944 and 1955 with the work of Paul Lazarsfeld, Bernard Berelson, Hazel Gaudet and later Elihu Katz.[3] *The People's Choice*, published in 1944, followed by *Personal Influence*, published in 1955, shifted the focus of analysis. Influence does not pass directly from the media to individuals. It passes through intermediaries, opinion leaders, primary groups and other forms of social mediation. This is what became known as the two-step flow model: the effect of a message depends on the social structure through which it travels. This finding remains fundamental in the cyber domain, as it explains why information that is highly visible may have little social impact, while less widely disseminated content may produce powerful effects if it passes through credible intermediaries or tightly knit communities.

This relational model was further supported by subsequent research on networks. The work of Duncan Watts, particularly in the early 2000s, demonstrated[4] that diffusion depends on network topology, adoption thresholds and the local properties of connections. In the 1990s and 2000s, Manuel Castells conceptualised the network society as the defining structure of contemporary societies. During the 2010s, Sinan Aral subsequently examined the spread of information online and the conditions under which content is disseminated. Cyberspace is a direct heir to this tradition: the digital environment is not a simple aggregation of exposed individuals, but a system of relationships, channels, communities, platforms and intermediaries.

The study by Bond et al., published in Nature in 2012, provides a major empirical reference point.[5] Examining 61 million Facebook users during the 2010 US elections, it demonstrated that social messages can influence political expression, information-seeking behaviour and actual voting decisions. Yet the main contribution of the study is not just the scale of the experiment. It is its demonstration of the role of social relationships: the

[3] Paul F. Lazarsfeld, Bernard Berelson and Hazel Gaudet, *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*, New York, Columbia University Press, 1944; Elihu Katz and Paul F. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*, Glencoe, Free Press, 1955.

[4] Duncan J. Watts, *Six Degrees: The Science of a Connected Age*, New York, W. W. Norton, 2003; Manuel Castells, *The Rise of the Network Society*, Oxford, Blackwell, 1996; Sinan Aral, *The Hype Machine*, New York, Currency, 2020.

[5] Robert M. Bond et al., 'A 61-million-person experiment in social influence and political mobilization', Nature, vol. 489, 2012, pp. 295-298.

effect does not result from abstract exposure, but from socially situated exposure, particularly through friends and friends of friends. For the cyber field, this means that a simple reach indicator is insufficient. What matters is the combination of exposure, credibility of the intermediary, position within the network and the target's behavioural proximity.

The third major scientific research phase focused on cognitive psychology and decision-making. During the 1970s and 1980s, Daniel Kahneman and Amos Tversky demonstrated[6] that individuals do not process information like rational computers. They rely on heuristics and are subject to anchoring, availability, framing and loss-aversion biases. A message therefore derives its impact not only from its content, but also from how it activates a cognitive shortcut. In 1986, Richard Petty and John Cacioppo proposed the Elaboration Likelihood Model, distinguishing between a central route to persuasion, which involves deep processing, and a peripheral route, which relies on contextual cues. In 1984, in his book entitled *Influence: The Psychology of Persuasion*, Robert Cialdini identified a number of robust principles, including social proof, authority, scarcity, commitment and reciprocity, which continue to underpin persuasive communication today.

This work settled one essential question: influence is not the same as exposure. An individual may see a message without processing it, process it without believing it, believe it without acting upon it, and cyber environments amplify this dissociation. Users are exposed to massive and often contradictory information flows while their attention is fragmented. Information overload, short-form content, the speed of dissemination and popularity signals all alter the conditions under which information is processed. This makes the number of views a very weak indicator. It measures potential access to the stimulus, but not its conversion into a decision.

[6] Daniel Kahneman and Amos Tversky, 'Judgment under Uncertainty: Heuristics and Biases', Science, vol. 185, no 4157, 1974, pp. 1124-1131; Daniel Kahneman and Amos Tversky, 'Prospect Theory: An Analysis of Decision under Risk', Econometrica, vol. 47, no 2, 1979, pp. 263-291; Richard E. Petty and John T. Cacioppo, 'Communication and Persuasion: Central and Peripheral Routes to Attitude Change', New York, Springer, 1986; Robert B. Cialdini, *Influence: The Psychology of Persuasion*, New York, William Morrow, 1984.

The fourth major phase examined discursive frames and narratives. Erving Goffman published *Frame Analysis* in 1974.[7] During the 1980s and 1990s, George Lakoff developed work on conceptual metaphors and political framing. Teun van Dijk further advanced critical discourse analysis by linking linguistic structures, social cognition and power relations. This body of research has long demonstrated that language is not merely a superficial layer. It shapes perceptions, activates interpretative frameworks, assigns responsibility, ranks threats and stabilises moral oppositions. Contemporary narrative-centred approaches are therefore not breaking entirely new ground. Rather, they are adapting a long-established insight to digital environments: messages do not operate as isolated units, but as narrative and interpretative configurations.

This reminder is important for the French debate. Recent approaches that emphasise narratives, serial patterns of discourse or narrative structuring make a valuable contribution to the analysis of information environments. They make it possible to track continuities, repetitions, recurring motifs, collective actors, patterns of accusation and sequences of victimisation or delegitimisation more effectively. They are relevant for understanding what is circulating, but they do not solve the problem of measuring effects. A coherent narrative is not necessarily influential, and a visible narrative does not necessarily lead to behavioural change. A marginal narrative can produce an effect if it reaches actors who are close to an action threshold. Conversely, a widely disseminated narrative may never become more than background noise.

The fifth major phase focuses on the empirical measurement of effects, and its conclusions are far less encouraging than is often assumed. The meta-analysis by Joshua Kalla and David Broockman, published in the *American Political Science Review* in 2018,[8] examined the effects of campaign contact and political advertising on electoral choices. Their conclusion is clear: in US general elections, the best average estimate of persuasive effects on candidate choice is close to zero. Their review includes a systematic meta-

[7] Erving Goffman, *Frame Analysis: An Essay on the Organization of Experience*, Cambridge, Harvard University Press, 1974; George Lakoff and Mark Johnson, *Metaphors We Live By*, Chicago, University of Chicago Press, 1980; Teun A. van Dijk, *Discourse and Power*, Basingstoke, Palgrave Macmillan, 2008.

[8] Joshua L. Kalla and David E. Broockman, 'The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments', American Political Science Review, vol. 112, no 1, 2018, pp. 148-166.

analysis of 49 field experiments. This result does not mean that influence is impossible. It means that the existence of an effect cannot be assumed solely on the basis of exposure. In saturated, polarised, identity-driven or already highly structured environments, an additional message may have no measurable impact.

This finding is of critical importance for digital influence and the cyber field because it challenges the assumption that more messages, greater visibility or faster dissemination automatically produce greater effects. Increased volume could strengthen a campaign, but it could also lead to overload, trivialisation, rejection or even benefit an opposing narrative. Cyber logic often encourages swift reactions, because the technical environment rewards detection and reaction. But influence is not malware: dealing with it is not simply a matter of identifying a threat and eliminating it. An information response may be effective, of no use or counterproductive depending on the audience, the channel, the timing, the credibility of the source and the proximity of individuals to a behavioural threshold.

This is where the problem of measurement becomes central. There is now a substantial body of knowledge in the field: networks matter, social intermediaries matter, cognitive biases matter, discursive frames matter, and platform signals matter. Yet these elements alone are insufficient for determining whether an operation has produced influence.

Strictly speaking, the effect of influence should be understood as behavioural change relative to a situation in which no intervention occurred. This is precisely the crux of the model advanced here: observable variables (production, visibility, circulation, engagement, coordination and narrative recurrence) describe the information environment, but do not in themselves measure influence. What should be measured is counterfactual behavioural change, in other words, the difference between what a population does when exposed to an intervention and what it would probably have done in the absence of that intervention.

This is not to suggest that narrative approaches, engineering methods or cognitive resilience programmes are of no use. More precisely, they currently have a prominent place in the debate because they focus on objects that are visible, institutionally convenient and scientifically tractable. Narratives can be described, networks can be mapped and mechanisms can be designed. Resilience can be tested through questionnaires or experiments. But the key

issue, namely the behaviour that would have occurred in the absence of intervention, often remains absent because it is the most difficult to assess.

This difficulty is even greater in the cyber domain, where operations are hybrid, audiences fragmented, data incomplete, platforms unstable, effects delayed and adversaries adaptive. A campaign may seek to encourage sharing, but it may also aim to create doubt, discourage, disorganise, divert attention, generate fatigue, undermine confidence in a procedure, make a decision more costly or render a future course of action acceptable. Many of these effects are not directly reflected in platform metrics. They emerge through weak, distributed and sometimes delayed behaviours.

The scientific implication is clear: the next frontier does not lie in producing yet another typology of narratives or a more sophisticated dashboard. It lies in linking the available indicators to a theory of effects. Such a theory must distinguish between exposure, susceptibility, cognition, the social norm, the cost of action, feasibility and behaviour. It must also compare scenarios: a public response, a discreet response, inoculation, correction, silence, a channel change or the mobilisation of local intermediaries. Without comparisons, evaluation becomes little more than a hindsight narrative.

The first conclusion is therefore as follows: the field of digital influence within the cyber domain does not suffer from a lack of data, nor even from a lack of concepts. What it lacks is a measurement architecture that assigns each concept its proper place. Narratives describe structures of meaning, networks describe circulation, engagement metrics describe interaction, and cognitive approaches describe reception. But influence, in the strict sense, begins only when behaviour is altered relative to what would have occurred in the absence of intervention.

II. THE LIMITATIONS OF NARRATIVE ANALYSIS, COGNITIVE ENGINEERING AND RECEIVER RESILIENCE IN MEASURING INFLUENCE

The assessment made in the AID report in 2018 provides a clearer lens through which to examine the approaches that currently dominate the French debate on digital influence. Since then, the field has not narrowed; on the contrary, it has become richer. France now has a National Strategy for Countering Foreign Information Manipulation 2026-2030, published on 11 February 2026, structured around four priorities: strengthening the nation's resilience, regulating platforms and generative AI, consolidating detection-attribution-response capabilities, and enhancing European and international cooperation. This strategy confirms that digital influence is now being treated as an issue of national security, where the cyber domain, the protection of public debate, Open Source Intelligence (OSINT), platform regulation and collective resilience intersect. It also formalises an important reality: combating information manipulation is no longer solely a matter of communication, but is now part of an interministerial system for protecting the information space.

This increased institutional commitment provides a framework for research and for practices that were previously dispersed, but it does not resolve the problem of measurement; it merely shifts it elsewhere. Three approaches currently occupy a disproportionate place in the debate: narrative analysis, the cognitive engineering of mechanisms, and the cognitive resilience of receivers. All three approaches are useful; none should be dismissed outright. Their weakness is not that they are wrong, but that they are situated upstream of, or alongside, the object that should be measured. They describe structures, design interventions or strengthen individual capacities, but on their own they do not yet make it possible to estimate the behavioural effect of an influence intervention in a cyber environment.

The first approach is centred on narrative. It consists in viewing information manipulation not as an accumulation of false claims, but as repeated,

stabilised and cumulative narrative structures. The approach is intellectually sound, because it extends longstanding work on framing, narrativity and discursive structures. In the recent French debate, this approach has been explicitly articulated through the idea of understanding disinformation as a 'serial narrative', that is, as a discursive construction that develops through episodes, characters, recurring conflicts and continuous interpretative frameworks. In an article published in *Le Rubicon* in November 2024, Paul Charon advocates this approach, presented as a literary approach to information manipulation.[9]

The contribution of this approach is obvious: it corrects the naïve assumption that each false or manipulative piece of content can be treated as an isolated object. An influence operation rarely works through a single message. It works through repetition, familiarisation and the stabilisation of an interpretative world. In the cyber field, this insight is useful: following a data breach, a ransomware attack, the compromise of an institution's systems or an act of digital sabotage, the narrative surrounding the event can establish a lasting interpretative framework. It can turn a technical vulnerability into evidence of political incompetence, an opportunistic attack into a strategic humiliation, or a local disruption into a sign of systemic collapse. Narrative analysis therefore helps to understand how a cyber sequence becomes an information sequence.

However, this approach does not measure influence: it describes the structure of discourse and does not tell us whether that discourse has altered behaviour. In other words, it allows us to determine that a narrative is coherent, repeated, connected to earlier episodes and that it constructs a moral universe, but it does not allow us to determine whether a population has changed its actual behaviour: whether or not it cooperates with an institution, shares information, follows cybersecurity instructions, migrates to another channel, participates in a mobilisation effort, or places its trust in an authority. In a cyber environment, this distinction is crucial. A hostile narrative can be highly structured and remain confined to audiences already convinced. It may be spectacular and have no behavioural consequences. Conversely, a weak narrative signal can produce a significant effect if it reaches a segment already close to an action threshold.

[9] Paul Charon, '*Lire la désinformation comme un récit sériel : pour une approche littéraire des manipulations de l'information*', Le Rubicon, 13 November 2024.

The problem, therefore, is not narrative analysis as such. The problem lies in the implicit assumption that it should occupy a central place in evaluation. It enriches the analysis of the information environment, but it is not a measurement of effects. It is based on an input variable: the narrative. The scientific object that should be measured is an output variable: altered behaviour. Between the two there is a missing chain that links exposure, credibility, perceived norms, cost, feasibility, identity compatibility and action. This is precisely what narrative dashboards do not provide. They can tell us that a narrative is gaining traction. However, without an additional model, they cannot tell us whether that traction changes what individuals actually do.

The second approach is cognitive engineering and mechanism design. It is represented by research work that proposes shifting from a logic of observation to a logic of design: developing a fine-grained understanding of audiences, prototyping responses, experimenting, adjusting, and integrating feedback. In an article published in February 2024, entitled '*Un design lab. pour la sécurité cognitive'*,[10] Axel Ducourneau argues for a permanent mechanism based on six principles: actor-centred analysis, a forward-looking approach, horizontal and vertical coherence, experimental agility, speed of execution, and the iterative integration of results through prototyping. The article also insists on an 'emic' understanding of target populations, that is, an understanding grounded in their own concepts and systems of thought.

This approach corrects a common weakness in influence mechanisms: their tendency to produce messages from the centre, with an insufficient understanding of target audiences. In the cyber field, this weakness is common. A government administration, military organisation, company or platform may respond to an information crisis using categories that make sense internally but do not correspond to how audiences perceive the situation. A cybersecurity alert may be technically accurate yet socially inaudible. An institutional response may be appropriate to the sender and counterproductive for the receiver. A correction can reassure an already trusting community but deepen distrust within a sceptical one. An actor-centred approach is therefore essential.

[10] Axel Ducourneau, 'Un design lab. pour la sécurité cognitive', Ingénierie cognitique, vol. 7, no 1, 2024, pp. 88-93, DOI: 10.21494/ISTE.OP.2024.1094.

However, cognitive engineering faces a similar limitation to the narrative approach. It improves the design of interventions, but it does not necessarily provide the criterion required to establish that an intervention has produced an effect. Testing a mechanism in a given context does not automatically indicate what would have happened in its absence. Without a baseline, without the ability to compare different options, without an estimation of net effect and without accounting for amplification risks, tests can become a well-organised exploration rather than an evaluation.

This limitation is particularly pronounced in cyber environments. An information operation tied to a cyber incident unfolds within an unstable context: official announcements, secondary leaks, expert commentary, journalistic interpretations, adversary reactions, technical rumours, user anxiety, legal constraints, and the timescale for remediation. If a response is implemented and information noise subsequently decreases, it is tempting to attribute that decrease to the intervention. Yet it could also result from the natural exhaustion of the news cycle, the emergence of another issue, the correction of a technical problem, account closures, a platform decision or a change of strategy by an adversary. Without a counterfactual estimate, effectiveness remains interpreted rather than measured.

Cognitive engineering must therefore be repositioned within the evaluation chain. It is useful when designing actions, but insufficient when it comes to measuring effects. It produces action hypotheses, prototypes and improvement loops, but without a causal architecture it cannot determine which of several options is preferable: a public response, a discreet response, no response, the use of a local intermediary, prior inoculation, a change of channel, or technical action targeting the dissemination infrastructure. Otherwise, the risk is that activity will be confused with effectiveness: because a mechanism is agile, fast, contextualised and iterative, it is assumed to be relevant. Yet in digital influence, a well-designed action can be of no use, a locally useful action can be harmful overall, and a silent action may be superior to a visible response.

The third approach shifts the analysis towards the receiver by focusing on cognitive capacities such as discernment, resistance to bias and the ability to identify manipulation techniques. This approach is part of a much older tradition than is sometimes acknowledged.

As early as the 1960s, the work of William J. McGuire introduced[11] the concept of psychological inoculation: exposing individuals to weakened versions of manipulative arguments could strengthen their future resistance. This idea has since been widely taken up, tested and debated, particularly in the fields of social psychology and persuasive communication.

Recent research makes it possible to be more specific. Interventions such as *Bad News* or *Go Viral!* expose participants to weakened versions of manipulation techniques (impersonation, polarisation, emotional appeals, conspiracy theories and source discrediting) and then measure their ability to recognise these techniques in experimental content.[12] Discernment can therefore improve, but the result remains situated. It primarily captures a capacity for judgement measured under controlled conditions, rather than a lasting change in behaviours such as sharing content, following a security instruction, refusing to click, migrating to a reliable channel or placing trust in an institution during a crisis.[13]

However, this approach is neither new nor free from limitations, and it has been the subject of recurring criticism.

First, the literature has long emphasised that the effects of inoculation remain contextual and dependent on exposure conditions. The work of Sander van der Linden and other researchers shows that these effects can diminish over time, vary across audiences and depend heavily on the presentation format and its context. Inoculation is not a universal mechanism; it is a situated intervention.

Second, several studies in psychology and political science have identified a persistent disconnect between improved judgement and behavioural

[11] William J. McGuire, 'Inducing Resistance to Persuasion: Some Contemporary Approaches', in Leonard Berkowitz (dir.), Advances in Experimental Social Psychology, vol. 1, New York, Academic Press, 1964, pp. 191-229.

[12] John A. Banas and Stephen A. Rains, 'A Meta-Analysis of Research on Inoculation Theory', Communication Monographs, vol. 77, no 3, 2010, pp. 281-311; Cecilie S. Traberg, Jon Roozenbeek and Sander van der Linden, 'Psychological Inoculation against Misinformation: Current Evidence and Future Directions', The ANNALS of the American Academy of Political and Social Science, vol. 700, no 1, 2022, pp. 136-151.

[13] Jon Roozenbeek and Sander van der Linden, 'Fake News Game Confers Psychological Resistance against Online Misinformation', Palgrave Communications, vol. 5, art. 65, 2019; Melisa Basol, Jon Roozenbeek and Sander van der Linden, 'Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity against Fake News', Journal of Cognition, vol. 3, no 1, 2020.

change. Individuals may improve their ability to detect manipulation without this leading to changes in their behaviour. This limitation is consistent with broader findings, including those of Joshua Kalla and David Broockman, which show that persuasive effects are often weak in saturated environments.

Third, this approach tends to isolate cognition from its social environment. Yet, as Cristina Bicchieri and research on social norms and collective behaviour have shown,[14] decisions depend as much on what individuals believe as on what they think others are doing or expect. The ability to identify manipulation is not enough to bring about different behaviour if perceived norms, costs or constraints remain unchanged.

This limitation is particularly evident in digital and cyber environments. Users may be educated about risks and capable of recognising certain techniques (phishing, deepfakes, impersonation), yet still not change their behaviour. Time pressure, practical constraints, information overload and social dynamics continue to shape decision-making. In this sense, cognitive resilience is a favourable condition, but not a measure of influence. It affects an intermediate variable, namely the capacity for judgement, yet does not guarantee that action will be taken. Improvement in discernment is measurable; the effect on action much less so.

The three approaches currently shaping the French debate should therefore be ranked in order of importance. Narrative analysis makes it possible to describe structures of meaning. Cognitive engineering makes it possible to design more appropriate interventions. Cognitive resilience makes it possible to strengthen certain capacities of the receiver. Yet none of these approaches, on its own, constitutes a measure of influence. They deal with discourse, the mechanism and the receiver respectively. The missing object remains altered behaviour.

This ranking is not an external criticism; it stems from the very nature of the problem these approaches seek to address. If digital influence is an effect, then that effect must be defined. If the effect is behavioural, then a target behaviour must be identified. If that behaviour can be influenced by multiple factors, then it is necessary to estimate what would have happened in the

[14] Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, Oxford, Oxford University Press, 2016.

absence of intervention. And if that situation without intervention cannot be observed, then a counterfactual approximation must be constructed. The model presented in the reference article formulates this point precisely: observable variables describe the environment, but influence must be estimated as behavioural change relative to a scenario without intervention.

The French debate today tends to be organised around objects that are relevant but incomplete. Narratives attract attention because they are accessible, interpretable and intellectually rich. The design lab attracts attention because it gives the impression of action, agility and operational modernity. Cognitive sovereignty attracts attention because it offers a human, democratic and defensive response to information overload. All three are necessary components of a digital influence policy; none is sufficient on its own for scientifically evaluating influence.

What is required is to reformulate the problem in operational terms. In cybersecurity, a distinction is normally made between indicators of compromise, indicators of attack, business impacts, remediation and residual risk. A similar logic must be applied to influence. Information signals are indicators of activity, not necessarily indicators of impact. Narrative coherence is an indicator of structure, not necessarily an indicator of effects. Engagement is an indicator of interaction, not necessarily an indicator of transformation. Cognitive resilience is an indicator of capacity, not necessarily an indicator of behaviour.

This cyber analogy helps to clarify the crux of the problem. No one would seriously conflate the number of network packets with the business impact of an intrusion. Yet in digital influence, message volume and engagement levels are still often conflated with actual effects. It is this confusion of levels that must be corrected. The field needs a measurement doctrine that distinguishes between activity, exposure, reception, decision-making and behaviour. As long as these levels continue to be conflated, debates will continue to produce concepts that are useful but insufficient.

The second provisional conclusion is therefore as follows: the approaches that currently dominate the debate should not be abandoned; they should be subordinated to the measurement of effects. Narrative becomes an input variable, and engineering becomes a method for producing options. Resilience becomes a factor in the modification of cognitive states, but the central question remains: which option alters which behaviour, within which

population segment, with what risk of amplification, relative to inaction? It is this question that introduces the third and final part.

III. FROM CYBER OBSERVATION TO EFFECT ESTIMATION

The first two parts of this report lead to a precise observation: the field of digital influence lacks neither data, nor tools, nor theoretical traditions. It is capable of observing the circulation of content, analysing narratives, mapping networks, designing mechanisms and strengthening some of the receiver's cognitive capacities. The difficulty, therefore, lies not in determining whether digital influence exists, nor even through which channels it can circulate. The difficulty lies in determining how to evaluate it as an effect. Put more precisely, how does one determine that an information intervention in a cyber environment has altered behaviour relative to what would have occurred in its absence?

This is a familiar problem in cybersecurity, although it is rarely formulated in these terms within the field of influence. When an organisation evaluates a cyberattack, it does not simply measure technical activity. It distinguishes between indicators of exposure, indicators of compromise, operational impacts, remediation costs and residual risks. A spike in network traffic does not automatically translate into a business impact. An intrusion attempt is not automatically a compromise. A compromise does not automatically result in service disruption. The evaluation chain requires several levels to be connected: hostile activity, exploited vulnerability, technical effect, business effect, cost and decision.

Digital influence should be treated with the same rigour. A high volume of messages, high visibility and narrative coherence do not constitute an effect. Similarly, improved discernment does not in itself amount to a behavioural effect. These are signals, conditions, intermediate variables or risk factors. The effect lies further downstream: in an observable, or reasonably inferable, change in a target behaviour.

This distinction is essential in the cyber field, because contemporary operations frequently combine technical action with information action. A data breach may be intended to cause social disruption or a technical

compromise. A ransomware attack may be accompanied by a public pressure campaign targeting the victim. A hack-and-leak operation can be less about revealing information than about destroying trust in an institution. A phishing campaign can be supported by narratives invoking crisis, urgency or authority. Information manipulation can prepare the ground for a technical compromise by lowering user vigilance. In all these cases, the relevant effect is not merely informational. It is behavioural: clicking, sharing, yielding, paying, remaining silent, accusing, disobeying, disengaging, circumventing a procedure, losing trust or changing a security practice.

The scientific question must therefore be formulated precisely: how can the behavioural change produced by a digital influence intervention be estimated when data are fragmented, audiences heterogeneous, platforms unstable and control groups rarely available? This formulation calls for four clarifications.

The first clarification concerns the dependent variable. There can be no measurement of influence without a target behaviour. In many analyses, the object remains vague: influence, perception, support, exposure, manipulation and mobilisation are discussed as though they referred to the same phenomenon. Yet these are different levels. Influence can seek to prompt a digital action, such as sharing content, clicking a link, joining a channel, reporting an account, leaving a platform or changing a security setting. It can seek to prompt social action, such as participating in a demonstration, refusing an instruction, supporting a position, boycotting a brand or spreading a rumour within a private group. It can seek to prompt an institutional action, such as undermining trust in a public administration, influencing a public decision, slowing a procedure or making a response politically costly. As long as the target behaviour is not defined, measurement remains imprecise.

The second clarification concerns the causal chain. Between content and action there are several intermediary stages. Exposure means that an individual or group has encountered a piece of content. Attention means that they have actually processed it. Credibility means that they deemed it plausible or legitimate. Identity compatibility means that the content does not conflict with the individual's or group's affiliations or values. A perceived norm means that they believe other relevant people think or act in the same way. Feasibility means that they are able to act. Cost means that they consider the risk associated with the action to be acceptable. Action appears

only at the end of this chain, and sometimes only once a threshold has been crossed. A user can be exposed to a fake cyber alert without clicking on it; can believe it without sharing it; can share it without changing their practices; or can change their practices without that change being lasting. Measuring exposure instead of action is therefore tantamount to interrupting the chain at the wrong point.

The third clarification concerns audience heterogeneity. In a cyber environment, the population is never homogeneous. Segments differ in terms of their level of institutional trust, their digital literacy, their ideological proximity to a narrative, their exposure to platforms, their sensitivity to technical evidence, their professional role, their degree of vulnerability and their position within a network. The same intervention can produce opposite effects across different segments. A public response can reassure an already supportive audience, irritate a sceptical audience and give visibility to hostile content among undecided audiences. A technical correction may be understood by experts but cause confusion among non-specialists. An awareness campaign can increase vigilance among one section of the public yet produce attentional fatigue in another. The average effect is therefore of limited value if the distribution of audiences close to an action threshold is unknown.

The fourth clarification concerns the counterfactual. Observing a change after an intervention is not sufficient to establish that the intervention caused that change. In cyber environments, this difficulty is always present. A hostile campaign may peter out naturally because the news cycle runs its course. A public response may appear effective when the observed change actually results from the platform removing accounts in parallel. A rumour may continue to spread despite an effective response, because an exogenous event is fuelling it. A decline in engagement may indicate a loss of interest, but it may equally indicate a migration to closed channels. Without an estimate of what would have happened in the absence of intervention, attribution remains fragile.

These four clarifications shift the analytical framework from a logic of monitoring to a logic of evaluation. Monitoring answers the question: what is happening within the information environment? Evaluation answers a more demanding question: what difference does the intervention make? The two approaches are complementary, but they are not equivalent. An organisation may have excellent monitoring capabilities but only limited

evaluation capabilities. In fact, this is the most common situation: many signals, little causality.

To make progress, digital influence must be treated as a problem of estimation under constraints. In an ideal world, the effect of an intervention would be measured through a controlled experiment: an exposed group, an unexposed group, observed behaviours, and comparable conditions. In operational cyber settings, this is rarely available. Populations are not randomly assigned, adversarial campaigns unfold in real time, platforms evolve, data are incomplete, important behaviours sometimes move to closed spaces, and decisions must be made before all the information is available. It is therefore necessary to rely on imperfect but explicit estimates.

This is where the concept of a synthetic population, or synthetic counterfactual, becomes important.[15] The aim is not to reproduce a real population perfectly. Rather, it is to construct a disciplined approximation of what would probably have occurred in the absence of intervention, and then compare that trajectory with different intervention scenarios. This approach is already familiar in other fields: epidemiological forecasting, crisis simulation, operational planning, economic modelling and public policy evaluation. It becomes necessary in digital influence as soon as the object being measured is a causal effect and experimental control is absent. In such circumstances, a model capable of meeting this requirement is needed. Indeed, when experimental or quasi-experimental controls are unavailable, a synthetic population becomes the minimum rational architecture for estimating behavioural effects under conditions of data scarcity. It does not predict minds; it specifies a baseline, compares scenarios and produces a decision bounded by uncertainty.

In a cyber environment, such an architecture would make it possible to compare multiple options. When faced with a hostile campaign following a data breach, should an organisation respond publicly, allow the narrative to die out, issue a technical response, mobilise independent intermediaries, target exposed audiences, act through platforms, or prepare a preventive inoculation? Each option modifies different variables. A public response increases visibility but can reduce uncertainty. A discreet response limits

[15] Alberto Abadie, Alexis Diamond and Jens Hainmueller, 'Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program', Journal of the American Statistical Association, vol. 105, no 490, 2010, pp. 493-505.

amplification but can fail to reach the most vulnerable audiences. A third-party intermediary increases credibility but reduces control over the message. Platform-based action reduces circulation but may fuel a narrative of censorship. Preventive inoculation improves future resistance but does not necessarily extinguish an immediate crisis. Non-response may be the best option if the narrative is weak and intervention risks drawing attention to it.

This type of comparison is impossible using engagement metrics alone. It requires an evaluation framework that distinguishes between activity, exposure, reception and behaviour. It also requires risks to be incorporated into the decision score. In digital influence, the useful effect is not simply the positive contribution of an action; it is what that action produces after deducting the risks of amplification, backlash, loss of credibility, saturation and displacement towards spaces that are more difficult to observe.

Amplification is probably the most underestimated risk. In the cyber field, an official response to a technical rumour can lend new legitimacy to that rumour. Refuting a marginal piece of content can introduce it into the wider debate. Publishing a denial can provide the adversary with talking points, screenshots, what appears to be evidence of concern, or a pretext for renewed attacks. A poorly calibrated response can turn a minor information attack into a media event. Measurement systems that put a premium on speed and visibility of response encourage this risk. They reward visible activity rather than net effects.

Backlash constitutes a second risk. An intervention can produce an opposite reaction among certain audiences, particularly when the sender is not perceived as a legitimate source. An institution may tell the truth and still lose credibility if the source is not accepted by the target audience. In a context of distrust, a correction can reinforce suspicion. In a cyber context, this dynamic is common: an institutional actor denying an allegation of a vulnerability or compromise can be perceived as both judge and interested party. An independent, local or technical intermediary may therefore be more effective, even if less visible.

Saturation is a third risk. Digital environments are already overloaded. Adding another message does not necessarily create clarity. It can create noise. As a result, audiences may disengage, ignore alerts, normalise risks or develop information fatigue. In cybersecurity, this phenomenon is well

known in the form of alert fatigue. It also exists in influence operations: too many warnings about manipulation can ultimately weaken discernment instead of strengthening it.

A fourth risk is loss of credibility. An imprecise, premature or overly broad response can permanently weaken the source. In a cyber environment, where technical facts may evolve, caution is necessary. A hasty attribution, excessive minimisation or incomplete communication can be exploited by an adversary. Effect measurement must therefore take into account not only immediate behaviour, but also the future credibility of the source.

These risks demonstrate why the scientific question cannot be reduced to detection. Detecting a campaign is necessary, but insufficient. It is necessary to determine whether it is behaviourally dangerous, for which segments, over what time horizon, and which response offers the best effect-to-risk ratio. This logic is similar to cyber risk management: the priority assigned to a detected vulnerability depends on its exploitability, exposure, business impact and the availability of a corrective measure. Similarly, the priority assigned to a detected hostile narrative depends on its credibility, dissemination, audience, proximity to an action threshold and the risk of amplifying it.

The evaluation of digital influence must therefore adopt a threshold-based approach. In many situations, an operation does not transform an entire population; it moves a segment that is already close to a particular behaviour. That segment may be small but decisive. In a cyber crisis, this may involve customers on the verge of leaving a service, employees likely to disregard an instruction, journalists hesitating between competing interpretative frameworks, public decision-makers under pressure, or users likely to share a false alert. The average exposure of the population is then not the primary indicator. What matters is the size of the audience segments close to the threshold and the capacity of an intervention to move them.

This logic explains why empirical findings on persuasion are often weak on average, yet significant in certain contexts. General campaigns reach large numbers of people, but they are often individuals far from the threshold. Social or normative interventions may reach fewer people, but have a greater impact on those who are already prepared to act. The study by Bond et al. demonstrates the importance of social relationships in mobilisation, while Kalla and Broockman demonstrate the weakness of average persuasive

effects in saturated political environments. These findings do not contradict each other. They indicate that effect depends on the position of individuals within a social and behavioural chain.
Research must therefore shift from a general question about the degree of influence exerted by a campaign to a structured question: 'Which segment, exposed by which source, under which perceived norm, with what action cost, is moved to adopt which behaviour, relative to which baseline?' This formulation may seem heavier, but it is scientifically more accurate. It prevents influence from being treated as an inherent property of a message. Instead, it places influence within a relationship between content, source, audience, context and behaviour.

It also makes it possible to properly reintegrate the three approaches discussed earlier. Narrative analysis provides information about content and interpretative frameworks. Cognitive engineering provides information about the design and adaptation of interventions. Cognitive resilience provides information about the receiver's capacity to process or neutralise certain stimuli. However, measuring effects requires these three dimensions to be linked to behaviour. This is not an opposition. It is an architecture.

For the cyber field, this architecture could be organised around five levels. The first level is the context: technical incident, political crisis, media event, exploited vulnerability or prior state of trust. The second level is the information environment: dominant narratives, sources, platforms, intermediaries, adversarial signals or coordination. The third level is the population: segments, levels of trust, exposure, digital skills or proximity to a target behaviour. The fourth level is the intervention: public response, discreet response, platform-based action, third-party intermediary, inoculation or silence. The fifth level is behaviour: digital action, social action, institutional action, security decision, change in trust, mobilisation or withdrawal.

Such an architecture does not provide certainty. It provides discipline. It requires assumptions to be made explicit, levels to be distinguished, uncertainty to be estimated and options to be compared. This already represents a major step forward compared with dashboards that accumulate signals without linking them to decision-making. In digital influence as in the cyber field, the objective is not omniscience; it is the reduction of decision-making error.

Such a framework will have to be validated gradually. This validation process can begin with retrospective analyses: taking information crises associated with cyber incidents, reconstructing the data available before decisions were made, simulating several options and then comparing the recommendations with the trajectories that were actually observed. It can then move to prospective exercises: producing estimates before intervention and assessing their calibration. Finally, expert judgement alone can be compared with expert judgement assisted by a model. The criterion will not be perfect prediction, but greater consistency, fewer amplification errors, better identification of at-risk segments, and the ability to justify a decision not to respond when non-response is the superior option.

That last capability is essential. A serious evaluation system must be capable of recommending inaction. If every detection triggers a response, the system is not an evaluation instrument; it becomes a mechanism for justifying activity. Yet in digital influence, inaction can be rational. It can avoid amplification, preserve credibility, allow a narrative to exhaust itself, or simply mean reserving a response for a more favourable time. The cyber field already operates according to this logic: not every alert triggers the same response; not every vulnerability justifies the same level of urgency; not every anomaly becomes a major incident. Influence must reach the same level of maturity.

The final scientific question can therefore be formulated as follows: how can a framework for estimating digital influence in a cyber environment be constructed that links observable information signals to target behaviours, incorporates audience heterogeneity, compares multiple response options and estimates what would have happened in the absence of intervention?

This question marks the next frontier for the field. It does not consist in producing a new typology of narratives. It does not consist in multiplying intervention prototypes. Nor does it consist solely in strengthening cognitive resilience. It consists in turning these building blocks into an evaluation system. Until that stage is reached, the field will remain rich in descriptions and poor in measurement. It will be able to describe what is circulating, but not what is changing. It will be able to detect campaigns, but not always assess the level of threat they pose. It will be able to produce responses, but not always establish that those responses were preferable to silence.

The issue is all the more important because cybersecurity is entering a phase in which the boundaries between technical systems, information and behaviour are becoming blurred. Attacks are no longer targeting systems alone; they are targeting decisions. They are exploiting technical vulnerabilities, but also cognitive, organisational and social vulnerabilities. Measuring digital influence therefore means protecting a decision-making chain. This is not merely a question of communication. It is a question of security.

IV. RECOMMENDATIONS: DEVELOPING A DOCTRINE FOR MEASURING DIGITAL INFLUENCE

First, define the target behaviour before measuring.
An influence operation should not be evaluated based on a generic visibility indicator. It must first be considered in relation to a target behaviour: clicking, sharing, abstaining, applying a security instruction, leaving a platform, joining a channel, trusting a source, exerting pressure on an institution or changing a professional practice. Without a target behaviour, metrics only describe information noise.

Second, separate indicators of activity, reception and impact.
Dashboards should distinguish between three levels. Activity indicators describe what is circulating: volumes, accounts, hashtags, timing and coordination. Reception indicators describe what is being processed: attention, credibility, memorisation and self-reported agreement. Impact indicators describe what changes: decisions, actions, inaction, withdrawal, mobilisation, cooperation or security practices. Conflating these levels leads to overestimating the effectiveness of visible campaigns.

Third, make narratives, design and resilience subordinate to effect evaluation.
Narrative analysis, cognitive engineering and receiver resilience should remain part of the evaluation chain, but each in their proper place. Narratives provide information about frameworks of meaning. Design produces intervention options. Cognitive resilience strengthens the capacity for judgement. None of these elements on their own demonstrate influence. They must be linked to an observed or estimated behavioural effect.

Fourth, construct a counterfactual baseline.
Every evaluation should include an estimate of what would probably have happened in the absence of intervention. This baseline can be imperfect, but it must be explicit: the natural development of the conversation, the expected media cycle, exposed audiences, usual intermediaries, the prior level of trust, and the probable trajectory of the adversarial narrative. A

response should only be considered effective if it performs better than this reference trajectory.

Fifth, compare multiple options, including silence.
Evaluation should not merely answer the question: should action be taken? It should compare multiple options: public response, discreet response, third-party intermediary, prior inoculation, technical correction, platform-based action, channel change or non-response. In certain situations, controlled silence may produce a better net effect than a visible denial, particularly when the risk of amplification exceeds the expected benefit.

Sixth, incorporate the risks of amplification, backlash and saturation.
The effectiveness score must incorporate the possible negative effects of an intervention: giving visibility to a marginal narrative, reinforcing pre-existing distrust, providing the adversary with material for renewed attacks, inducing audience fatigue or weakening the credibility of the source. An information action should therefore be evaluated in terms of net effect, not visible activity.

Seventh, prioritise segments close to an action threshold.
Efforts should focus on audiences capable of changing behaviour, not merely those most heavily exposed. A campaign may reach many individuals who are already convinced, without producing any effect. Conversely, a segment that is small but already close to a threshold, such as employees tempted to circumvent a procedure, users considering leaving a service, journalists seeking an interpretative framework or decision-makers under pressure, may prove decisive.

Eighth, organise evaluation around five levels.
The measurement doctrine should be structured around five levels: the crisis context, the information environment, the population, the intervention and the behaviour. This architecture requires explicit specification of what is observed, what is inferred, what is decided and what is ultimately changed. It also makes it possible to justify a proportionate response and to retain a record of the assumptions used.

Ninth, validate models progressively based on lessons learned.
Evaluation models should first be tested on retrospective cases and then through prospective exercises. Recommendations formulated before

decisions are made should be compared with the trajectories subsequently observed: reduction in amplification, improved identification of at-risk audiences, greater accuracy of assumptions, and the capacity to recommend non-response. The objective is not perfect prediction, but the reduction of decision-making error.

CONCLUSION

There are already many building blocks in the field of digital influence: networks, narratives, biases, intermediaries, metrics, experimental mechanisms and resilience programmes. What is still lacking is an architecture that allows them to be organised around effects. In the cyber field, this architecture must impose a strict distinction between visibility and impact, activity and effectiveness, narrative and behaviour, action and outcome. This requirement places the counterfactual back at the centre of analysis. It is not enough to know what happened after an intervention. What is needed is to estimate what would probably have happened in its absence, and then compare the available options while incorporating the risks of amplification, backlash, saturation and loss of credibility.

Only under these conditions can digital influence be treated as a cyber object in its own right: an object that is no longer merely monitored, commented on or mapped, but measured in terms of its probable effects, risks and alternative courses of action. True progress will not come from producing a more detailed map of flows. It will come from knowing whether that map leads somewhere.

SKEMA Publika
SKEMA Business School, Grand Paris Campus
5 Quai Marcel Dassault – CS 90067
92156 Suresnes Cedex, France

Tel.: +33 (0)1 71 13 39 32
Email: publika@skema.edu
Website: www.publika.skema.edu