Viewpoints

Assessing Assessments: How Useful is Predictive Intelligence?

By WO2 John Hetherington and Wing Commander Keith Dear

Biography: Warrant Officer Class 2 John Hetherington serves in the British Army's Intelligence Corps. Currently an Arabic and Pashto linguist, he has served in Northern Ireland, the former Yugoslavia, Iraq and Afghanistan, as an analyst in the combat and operational intelligence roles. He has a long standing interest in psychology and war, presenting on psychology's application to tactical planning and execution.

Wing Commander Keith Dear is an RAF intelligence officer, CAS Fellow, DPhil candidate at Oxford University's Department of Experimental Psychology, and Research Associate at Oxford's Changing Character of War Programme. His professional experience is in Intelligence, Surveillance and Reconnaissance (ISR) analysing human behaviours and systems. His current studies examine the effect of surveillance on behaviour.

Introduction

"Tell me what you know, tell me what you don't know, tell me what you think, always distinguish which is which"[1]

General Colin Powell, Chairman US Joint Chiefs of Staff

This article suggests that the failure to retrospectively evaluate the accuracy of British military intelligence predictions limits its effectiveness as an aid to commanders. It argues that it must begin to include accuracy of prediction as a metric in the assessment of the professional effectiveness of our individual analysts and our intelligence organisations and process as a whole. Doing so will provide the feedback needed to improve and may also help to more carefully delineate the limitations of prediction.

A central responsibility of military intelligence staff is to predict the future actions and intentions of enemy, neutral and allied parties. But how accurate and useful are these predictions? At present, the British military would be hard pressed to answer this essential question with any rigour. No mechanism looks at how often an analyst's or analytical team's predictions are right or wrong. A retrospective evaluation of British military intelligence assessments in order to judge their accuracy would be the first step in improving future performance. This is no more than would be expected in any other domain from flight safety to a patrol debrief. In the absence of any work examining how accurate British intelligence prediction has been, this article looks at US and Canadian efforts to understand the accuracy of their predictions. It focuses first on identifying practical barriers to accurate prediction, and then on the theoretical aspect of prediction – what can be successfully predicted, or what are the boundaries of prediction?

The term assessment is sometimes used to indicate an analytical curation of available intelligence and other information to give a general picture to a commander. In order to remove any ambiguity, this article will use the term assessment to mean the summarised intelligence picture delivered to a commander – it is a retrospective or current perspective on events. The term prediction indicates that portion of an assessment which is an analyst's estimate of future enemy or third-party activity. This article covers predictions pertaining to the tactical, operational and strategic levels largely without distinction as what is at issue here resolves to one or more intelligence personnel making a prediction about the future, and in this dimension the similarities between the levels of war outweigh the differences.

US-funded research (the so-called 'Superforecasters' project) examining the accuracy of those it employs to make intelligence[2] predictions found that external forecasters without access to classified information were 30% more accurate in their predictions than the professionals with access to classified materials.[3] It appears that the base prediction accuracy has still not been published. However, an earlier study found that the average expert was "no more accurate than a dart-throwing chimpanzee".[4] The best forecasters in the first round of the project, those that had contributed the most in the 30% victory were grouped together as a 'Superforecasters'

team. They beat American intelligence analysts in accuracy of prediction by 60% in the following round of the project and, after a further round of selection, by 78% in the third. There was clearly great room for improvement amongst the professional predictors in the US intelligence community.

In contrast to the US results, David Mandel's subsequent work with Canadian intelligence showed that Canadian analysts achieved very high accuracy in their predictions.[5] Mandel's work is thought to have been more accurate for the following reasons: the predictions were on shorter timescales (e.g. 6-12 months, rather than 12-36 months), the analysts were not given anonymity and they were not necessarily acknowledged experts in the theory of their subject.[6]

Unless the various UK intelligence agencies and staffs know empirically or rigorously how right, or how wrong, they are, it is very difficult to *systematically* improve the systems and processes that lead to those successes and failures. Former British military intelligence officer Sean Ryan[7] summarizes the current situation within UK military intelligence admirably, writing that "… success rarely influences official appraisals of intelligence professionals. It would be impossible, as no records compare the end result to the prediction. Formal 'lessons learned' processes limit themselves to broad-stroke impressions, generic issues and localised procedures. No metrics record the accuracy of intelligence analysis on an individual, team or organisational level. Rarely is the question asked: 'why did we get that wrong?' Intelligence analysts are structurally divorced from responsibility for the accuracy of their assessments."

It is critical to know how accurate these predictions are, if they are to be improved.

Preliminary Diagnosis

There are a number of practical and theoretical issues with how UK intelligence staffs currently approach prediction. This article deals with the practical first. It shows the need for systematic review of predictive accuracy. It discusses the importance of language, the difficulty of assessing the probability of singular events, the overweighting of confidence in military culture and the possibility that analysis is valued too highly over the more clerical aspects of intelligence: processing and exploitation.

Assessment and prediction are at the heart of British Military Intelligence training and practice. From the commencement of phase 2 training, through every exercise and operation, intelligence staffs are obliged to predict the future actions and intentions of enemy or neutral forces and parties. Whether working at the tactical or the strategic level, prediction is central to the intelligence analyst's role.

It matters that the British military does not know how good it is at prediction. Giving a prediction, whether in a written or oral brief, has the effect of creating a position which the intelligence staff, the command staff, and the operators, may all buy into. They will see future intelligence through the prism of the first assessment, being reluctant to give it up and,

through confirmation bias, search for evidence to support it. Inaccurate predictions result in poor decisions and blindness to other threats and opportunities.[8] The psychological effects are well understood: by letting an incorrect assessment into collective understanding, the ability to perceive reality and predict correctly is reduced. In the jargon, the prediction becomes the baseline while the 'public commitment' makes it hard to get it out of the collective brain.

Precision of Language

Clarity of language is critical. Tetlock's findings were unambiguous on this: increased precision in outcome was correlated with increased precision in prediction; loose language must be avoided. This problem is formally recognised in doctrine. JDP 2-00 para 343 provides us with an uncertainty yardstick which applies numerical values to probabilistic language.

Qualitative Term	Associated Probability Range
Remote *or* highly unlikely	Less than 10%
Improbable *or* unlikely	15-20%
Realistic probability	25-50%
Probable *or* likely	55-70%
Highly probable *or* highly likely	75-85%
Almost certain	More than 90%

Figure 3.7 from JDP 2-00 – Defence Intelligence Uncertainty Yardstick.

This precise approach is critical and must be taught and applied at all levels. In practice its routine application appears confined principally to the higher Strategic and Operational levels only.

When prediction is hedged around with 'may', 'could' or 'possibly' without such a clear framework the results can be devastating. Tetlock shows how this phenomenon contributed to a number of poor decisions. When military and political staffs discussed the possibility of a Soviet attack on Yugoslavia (1951) or the chances of the Bay of Pigs operation succeeding (1961), those present understood radically different numerical chances for the same words. UK Defence Research recommended the adoption of these measures in 2002 to allow Commanders to better evaluate the weighting of risk.[9] Retrospective analysis of predictive accuracy would provide feedback on how widely this direction has been followed.

There is some evidence that the British Military may not be quite as committed to precision in intelligence analysis as it might like to claim. For example, one of the few publicly available intelligence predictions in the UK is that of a terrorist attack. MI5/the Security Service tells us

that the UK threat level for international terrorism is currently at SEVERE, meaning an attack is highly likely. The threat level has been at SEVERE or SUBSTANTIAL for 10 years, since 2006.[10] The agency explains that SUBSTANTIAL means an attack is a strong possibility while SEVERE means an attack is highly likely. There is no read-across to the DI Yardstick for 'strong possibility' which leads us to suggest this is a linguistic fudge masquerading as a prediction, a public exercise in risk aversion. This view is reinforced by reading across the SEVERE probability to the Yardstick. If SEVERE does mean an attack is *highly likely*, which the DI Yardstick tells us is a 75-85% chance of being attacked, no timeframe for the UK threat level from terrorism is given, which is itself a significant problem. It makes it impossible to judge whether the prediction was accurate. If the predicted timeframe is taken to be 24 hours, then the UK has had a greater than 75% chance of experiencing a terrorist attack every day since 2014. This is plainly absurd given the evidence of how rare terrorist attacks are in the UK. If politics is driving the prediction this must be opposed. It misleads the public, acts as a poor guide to policy and indeed it might be argued that in overweighting the strength and effectiveness of terrorist groups it serves to amplify the fear, the terror, they seek to create. If the risk aversion lies within the intelligence agencies, it must be addressed. It is in such highly charged times that rigour in predictions is most important.

Mandel's analysis of Canadian intelligence showed that there exists within the group a relatively clear shared understanding between commanders and analysts of what the less precise verbal formulations mean. It is highly likely that this understanding exists in the British military too and also highly likely that it is increased over time amongst individual intelligence and command staffs. But if accuracy in prediction is to be assessed, more precision is needed in saying what is meant by giving a clear numeric probability against which the analyst can be examined. Clearly, if multiple predictions are made with 90% accuracy, they should be right 90% of the time. Examining this would help analysts to recalibrate their predictions for habitual over- or under-confidence, and enable organisations to see and correct systematic biases or failures. Furthermore, in large, complex international coalitions the shared understanding between commander and staff in one area is unlikely to extend to other commands and staffs across the coalition. Rigour becomes ever more important as complexity and scale increase.

Predicting singular events against a numerical probability is more complex than predicting multiple events by extending an existing pattern or trend. Suggesting that there is a 35% chance of an event occurring may help a commander plan the disposition of his forces, but if the event subsequently happens it can't be known if it was random chance or a bad prediction. However, the given probability can be used as part of an evaluation of the totality of an analyst's or section's predictions. Even for singular events: if 70% of predictions only come to pass 33% of the time, there is a systemic or individual problem with the approach taken (or an extraordinary run of bad luck – possible, but little harm is done by a close examination of process and procedure even when it is erroneously cued by statistical improbability). It is more helpful then for a commander to know how good his or her intelligence staff is in making predictions, than it is for him or her to know the probability assigned to the prediction of a singular event.

Analysts' Behaviour and Skills

It may be that the wrong habits are encouraged in our analysts. To quote *Superforecasting*: "... people equate confidence and competence, which makes the forecaster who says something has a middling probability of happening as less worthy of respect". As one study noted, people "took such judgements as indications the forecasters were either generally incompetent, ignorant of the facts in a given case, or lazy, unwilling to expend the effort required to gather information that would justify greater confidence". The plausibility of the analyst may be more likely to convince the listener or reader that they are right, but they are not necessarily more likely to actually be right. Pioneering behavioural psychologist Daniel Kahneman notes that "declarations of high confidence may just tell you that an individual has constructed a coherent story in his mind, not necessarily that the story is true".[11] This preference for expressing certainty rather than doubt – added to group-think – was a key element in the Iraq intelligence failure. At a lower level, Ryan notes that: "In OPINT exercises, students are primarily graded on their delivery and plausibility. Rarely is there a mark for accuracy. Many exercises are deliberately scripted without a right answer. The instilled effect is a focus on presentation over content, because presentation is the tangible metric on which individuals are graded. Intelligence operators are incentivised to become salespeople rather than analysts, judged not by performance but by plausibility". Prediction should nearly always be tentative so that both analyst and commander keep the right mind-set about such intelligence predictions: they are often assessments of very difficult and dynamic situations based on incomplete information and error, and the unexpected should be routinely expected.

Finally, it is possible that prediction is over-emphasized in intelligence circles because it is a difficult and perhaps even "illusory skill". Kahneman uses this term in connection with stock market traders, describing them as highly skilled and hard-working professionals using every scrap of information to improve their predictions but still unable to consistently beat or match the stock market. This is because despite the high-level skills being employed, the task is usually impossible. If intelligence prediction is, indeed, an illusory skill then its importance is systemically over-rated in comparison with the more clerical aspects of intelligence work such as collection and dissemination, which are seen as more pedestrian. A full systematic review of how accurate and useful our predictions are may teach that displaying accurate information clearly and quickly has far greater value to all staff, from command to field.

The Boundaries and Value of Prediction

Some important conceptual challenges to analysts' ability to predict must be considered. Are the answers to the questions that the UK asks its analysts to predict even theoretically routinely predictable? 'Routine' is important here: everyone will get lucky sometimes. The difficulties in successful prediction in military contexts are legion: incomplete and incorrect information, a dynamic situation receiving constant new inputs, a very large number of possible states for each actor, with many actors and parties involved, a wide range of possible outcomes, and complex interactions of friendly and enemy forces. This leads

to frequent 'black swan' events (possibly more common in lower-intensity warfare): rare, inherently unpredictable and retrospectively rationalised to make them *appear* predictable – the narrative fallacy. However, even in high intensity warfare there are parallels in the attempts to predict singular events, like the day of an invasion.[12]

Professor Jim Storr suggests that high-intensity combat consists of a large number of elements in (1) a large number of different states (attacking/defending/delaying/regrouping etc); in (2) a large number of different spatial positions, which (3) change often in time, while (4) continuously interacting with other friendly and enemy elements dynamically and lethally. All this occurs in an environment in which decisions (i.e. future intentions) will be made in conditions of great – sometimes mortal – danger. In addition to all of this, each side will have imperfect information about both itself and the enemy. Although describing high-intensity combat, all the situations and levels of combat that our intelligence staffs must assess will share these attributes to a large degree. In these circumstances, which Storr categorises as "unutterably complex and [which] do not appear to be heavily determined", some might argue prediction is impossible. It is certain that it is very difficult.

Military intelligence staff training tends to be scenario based. In such an approach a 'GENFORCE mentality' can creep in: this is where an exercise enemy is used, which has stereotyped tactics and operates in an unrealistic 'zero-friction' environment. Accurate 'prediction' in these circumstances is analogous to solving a puzzle: finding the key bit of information will unlock the solution. Unfortunately, this tells us little about the ability to predict a real enemy operating in a dynamic and uncertain environment.[13]

Lieutenant Andy Mellows' excellent paper published in the Int Corps Journal *Cognito* examining intelligence and decision making contains a suggestion for improving scenario-based training. Mellows suggested that exercises are created which give real situations from the past and then testing intelligence staffs' ability to predict the outcomes, measuring their performance against actual outcomes.[14] Such a thoughtful approach to improve and further professionalising intelligence assessment and prediction is precisely what is needed and we could provide part of the answer. However, if Professor Storr is correct, caution is necessary in how this idea is implemented. Even the attempt at retrospective prediction based on clear historical example may fall victim to the narrative fallacy – assuming the outcome observed was the most probable, or even inevitable.

If the limits of what *should* be predicted are unknown, and the ability of our analysts and organisations to make accurate predictions is similarly unclear, the value of intelligence predictions is equally uncertain.

As a second order effect, this uncertainty calls into question the usefulness of IPB/IPE and doctrinal templates and similar products in real warfare. Professor Storr suggests that such products are rarely updated or even referred to after initial planning, so there is no way of

knowing if they aided accurate prediction.[15] Even more damningly, Major General Michael Flynn, senior military intelligence officer in Afghanistan, said in 2009: "The intelligence community's … culture is strangely oblivious of how little its analytical products, as they now exist, actually influence commanders. It is also a culture that is emphatic about secrecy but regrettably less concerned about mission effectiveness".[16]

With such enormous difficulties the intelligence community must do everything it can to improve accuracy in prediction. It cannot afford not to.

This article has attempted to diagnose some of the practical and theoretical problems that inhibit or prevent accurate intelligence prediction. It will briefly examine some objections to the diagnosis, answering them before turning finally to some recommendations.

Objections

Certain objections could be made to various aspects of the diagnosis. On the more practical side it might be argued that there is (1) no discontent with current intelligence practices and therefore no problem; or (2) that there is a shared understanding at all levels regarding the language used, both in the operations and intelligence staff making and using predictions, and with the audience for these predictions. Some may argue that (3) confident delivery is not the same as the analyst's confidence in the prediction, and does not lead deterministically to a commander's confidence in the prediction. All of this could be determined by a systematic retrospective analysis of intelligence predictions and the command and tactical decisions based on them, or made in contradiction of them.

Others might suggest that command and intelligence staffs are both happy to deal in uncertainty and neither is unduly concerned to over-defend stated positions. Again, an audit would reveal how true this. A brief acquaintance with psychology would suggest that it is unlikely to be true.

Some have suggested that there is not time to be more precise in making intelligence predictions and/or there is insufficient time to go back and analyse their accuracy. It was noteworthy that neither Tetlock nor Mandel suggest more accurate predictions take longer to make but in fairness it must be considered that the conditions in which the predictions were made were different. It is possible that the extra linguistic and probabilistic precision requires a change of habit rather than a longer time in formulation of the prediction. Making sufficient time for the retrospective analysis of predictive accuracy may be more difficult. Staffs may genuinely not have the capacity to do this. It may be advisable therefore for an outside agency or a different staff to carry out the analysis.

A notable challenge is that a prediction can be confounded by being right. That is, an analyst may be thought to have made an incorrect prediction when the action was in fact prevented by friendly activity, perhaps activity directly cued by the commander's desire to prevent the

predicted outcome occurring. The difficulty of friendly operations changing future enemy activity and so invalidating earlier predictions is clearly a valid objection, to a degree, however, this could be incorporated into the evaluation of the prediction e.g. was any action taken based on the prediction? Surely one of the most powerful indicators of the utility of our predictions is how frequently they cue action to forestall the predicted outcome. Thus this might be a key metric of the utility of the predictive intelligence. Again, it is a metric that is not currently tracked. Furthermore, any action taken should yield further information to confirm or refute the prediction's accuracy – again providing feedback to allow the analyst to improve in future.

This article notes a final practical objection: some suggest the last decade of war has redressed the balance between assessment and the other parts of the intelligence cycle. While a theoretical objection might argue that given sufficient current and doctrinal information, and sufficient analytical power, complex tactical movements can be predicted.[17]

The truth is, that most of these objections are impossible to completely refute and are equally difficult to uphold – the answers are unknown and are worthy of discovery through examination of our predictive accuracy. At best, commanders may gain greater confidence in their intelligence staffs; at worst, it may be discovered that some working assumptions are fundamentally unsound.

How do intelligence predictions improve?

Some recommendations arise from this article's analysis. First and foremost, **predictions must be subject to systematic, continuous review both on a collective and individual level**. Such a review might be most effective if impartial and thus should be led by a neutral third party – another staff, a commercial organisation or an academic partnership. The UK needs to know how good its intelligence staffs are in order to adapt its training and its approach, and analysts need to know how good they are as individuals in order to improve their individual performance.

To do this, the Yardstick must be applied universally, and the UK must begin to look at the accuracy of intelligence organisations, sections or individual analysts' predictive performance as part of such an assessment of their professional effectiveness. The simple measure of recording and then subsequently verifying the accuracy of each analyst's, section's and organisation's predictions should begin to bring a number of 'quick wins'. For example (1) the accuracy of the forecasts of the intelligence staffs will be known, in itself a useful aid to planning; (2) once they are known there will be a measurable incentive to improve the accuracy of the forecasts, (3) this baselining will create an incentive to minimise over-confident predictions and the subsequent damage they do to the commander's mental picture of events, and (4) Command staff may become happier with less certainty in their intelligence staffs' assessments and will gain greater confidence in analysts based on accurate results, not plausibility of presentation. Additionally, (5) intelligence and command staff will be better

able to understand broadly what can be predicted and to what level of accuracy while certain things may be ruled out as definitively unpredictable – in the jargon, the boundary conditions for successful prediction in military contexts will be discovered. Finally (6) the utility of certain staff procedures can be tested by how much or how little they contribute to the accuracy of subsequent predictions.[18]

The Operational Record from Ops TELIC, HERRICK, ELLAMY and perhaps even SHADER, as currently stored in electronic format, could readily enable an analysis of the accuracy or otherwise of recent intelligence predictions at given levels of detail and time period. It might also enable the intelligence community to understand the differences in outcome when predicting events at the Tactical, Operational and Strategic levels. It must be independent and would be well suited to a military-academic partnership.

The successful approach of Tetlock and the Intelligence Advanced Research Projects Activity (IARPA) programme could be adopted by identifying who the British 'Superforecasters' are and grouping them together, creating a sort of 'special forces' of forecasters.

Some recommendations emerge from the studies of Canadian and US intelligence which the British military would be wise to adopt. While Tetlock's research found most professional forecasters to be unreliable, he did find certain people who reliably made accurate predictions and identified the personal characteristics they possessed. These may be listed for simplicity:

- They made specific measurable forecasts.
- They constantly adjusted those forecasts in the light of new information. This information might only shift their prediction a few percentage points (the best forecasters were the most granular).
- They were not emotionally tied to their predictions.
- They were not adherents to a big over-arching political ideology which explains everything.
- They broke down problems into constituent elements.
- They were comfortable with basic mathematics.[19]
- They improved as forecasters as they gained experience.

These might form the spine of any future analytical training and/or assessment of individual or collective competence.

In response to the conceptual challenges to prediction, its conditions must be made easier by reducing the number of elements and thus increasing the degree of accuracy in space and time. Broad patterns will generally be a lot easier to predict than anything which requires detail and precision, and Storr[20] says that this is what expert military decision makers, i.e. good commanders, actually do.

Summary

This article suggests that the UK must seek to replicate, expand and institutionalise nascent efforts in the US and Canada to improve the accuracy of intelligence staffs' predictions. The first task must be to establish how good intelligence staffs are at making predictions. The second must be to make adjustments to training, processes and procedures based on what we have learned. Tetlock and Mandel have provided an invaluable pointer to where we need to go. It's up to the intelligence community to take the actions to get there.

Notes

[1] General Colin Powell, Chairman Joint Chiefs of Staff, "Intelligence Reform" 2004, http://fas.org/irp/congress/2004_hr/091304powell.html.

[2] These were strategic/geo-political assessments.

[3] David Ignatius "More Chatter Than Needed" 2013, https://www.washingtonpost.com/opinions/david-ignatius-more-chatter-than-needed/2013/11/01/1194a984-425a-11e3-a624-41d661b0bb78_story.html.

[4] And in the case of media pundits, worse...

[5] David Mandel, *Accuracy of Intelligence Forecasts From The Intelligence Consumer's Perspective*, Defence Research and Development Canada 2015.

[6] Tetlock & Mellers, *Judging Political Judgement*, Proceedings of the National Academy of Sciences of the United States of America July 2014.

[7] Sean Ryan, "Finding the Right Answer," *The RUSI Journal* vol.160 iss.4 (2015); 50-58.

[8] Storr, *The Human Face of War*. Professor Storr, former soldier and military scholar mentions an example of a J2 cell literally making up an attack and the authors have personally witnessed similar phenomena several times.

[9] Desimone and Charles, "Towards an Ontology for Intelligence Analysis & Collection Management", Qinetiq 2002 https://www.aiai.ed.ac.uk/project/ksco/ksco-2002/pdf-parts (paper 11).

[10] https://www.mi5.gov.uk/threat-levels.

[11] Daniel Kahnemann, *Thinking Fast and Slow*.

[12] There is a good example from the Yom Kippur War: http://gladwell.com/connecting-the-dots/.

[13] Ryan, "Finding the Right Answer", is similarly sceptical.

[14] Mellows, *Embracing intuition: how an understanding of Naturalistic Decision Making research can improve the provision of tactical intelligence*, 2015 Cognito.

[15] Storr, *The Human Face of War*, identifies targeting Boards, IPB and synchronization matrices as falling into this category.

[16] Flynn, Pottinger & Batchelor, *Fixing Intel: A Blueprint for Making Intelligence Relevant in Afghanistan*, (Voices From the Field Series, Center for a New American Security, 2010) quoted in Ryan, "Finding the Right Answer".

[17] The difficulties encountered in more general political, economic, climatic and even sports forecasting invites a certain scepticism here.

[18] Tetlock, *Superforecasting* is useful for its scoring system for those interested.

[19] Mathematical awareness appears to have a basic function in preventing an analyst holding logically incompatible views
[20] op. cit.