VALIDITY AND LIMITS OF ALGORITHMIC DECISION-MAKING

Gábor Hámori

Browsing an online bookstore these days, it comes as no surprise if the site automatically offers us certain books that are likely to capture our interest. We receive SMS messages from our telephone service provider recommending us new products suited to our demands. Applying for a bank loan, our request is now automatically processed by the credit institution in a matter of minutes. The common feature of the above examples is that, in each case, a forecast of our expected behaviour has been made using algorithmic predictive methods. But can mathematical and statistical methods still be successfully applied if, for example, we want to establish a prognosis for the future evolution of Bordeaux wine prices, or the durability of a marriage? Or do we prefer in such instances to trust in experts with experience in the given field? In the following study, relying on published research into the topic, we present a brief overview of the factors which determine whether we should rely on algorithmic tools or trust the assessment of experts in a given decision-making or predictive situation.

JEL codes: C10, C30, D70, D81
Keywords: decision theory, cognitive distortion, prediction, statistical forecasting, expert assessment

ERRORS IN THINKING THAT INFLUENCE OUR DECISIONS

In order to be able to make a satisfactory evaluation of algorithmic decision methods based on the role they play in actual decisions, let us first take into consideration (without making an exhaustive list) the typical errors in thinking which we tend to make on a daily basis, even unconsciously, and which do not fail to impact even expert value judgements, thus influencing the quality of the relevant decisions and forecasts.

Of all the errors in thinking we make, the type with perhaps the greatest impact is known as the *confirmation bias,* whereby, prior to making a decision, we are inclined to evaluate the available information in a way that, if possible, avoids conflict with our professional beliefs and views of the problem concerned. As a consequence, we typically filter out information which does not conform to our world view, or attach significantly less weight to it than would truly be justified. As the famous quote by *Warren Buffett* has it: “What the human being is best at doing is interpreting all new information so that their prior conclusions remain intact." The greatest problem is that the confirmation bias is typically not a conscious error, so that generally we simply "do not hear" information that runs counter to our own hypotheses.

The *disregarding of probabilities that define a problem*, and within this the *zero-risk bias* (*Slovic*, 2000), often leads to erroneous decision-making. For example, lottery players and other gamblers, by entering a game, certainly disregard the true probabilities inherent in the game, and thereby contribute to the extraordinary profitability of the gambling industry. To understand the zero-risk bias, let us suppose we can choose between two projects (*Dobelli*, 2013). In one case, we can reduce the number of deaths caused by environmental pollution from 5% to 2%. With the help of the other project, under an identical set of conditions (costs, time horizon, etc), a 1% mortality risk can be completely eliminated. Most people prefer the second alternative, despite the fact that the first case brings a 3 percentage-point improvement; in other words, three times more than the investment that promises a zero risk. A simpler but more frequent instance of disregarding probabilities is the *base rate fallacy*. The following example makes it easy to understand what this actually means. Let us suppose that Peter is a thin man with spectacles who likes Bach. What is more likely, that Peter is *a*) a taxi driver, or *b)* a literature professor in Budapest? Most people would wrongly opt for the second answer. There are a great many more taxi drivers living in Hungary than literature professors in Budapest. Consequently it is far more likely that Peter is a taxi driver, even one who likes Bach. Those opting for the second answer disregarded the statistical base rate.

The so-called *prognosis illusion* (*Dobelli*, 2013) can primarily be observed in the case of long-term expert predictions of complex systems. Berkeley professor *Philip Tetlock* evaluated more than 80,000 predictions by 284 experts (*Tetlock*, 2006.) The study revealed that the predictions were scarcely more accurate than if prognoses for the future had been generated at random. One surprising finding of the research was that it was forecasts by experts enjoying the media spotlight which tended to prove the most unreliable. With a hint of sarcasm, Harvard economist John Kenneth Galbraith summarised the phenomenon thus: "There are two kinds of forecasters: those who don't know, and those who don't know they don't know."

And finally we should mention *overconfidence*, where we typically overestimate our own knowledge level. Surprisingly, excessive self-confidence is a feature of experts – and, among them, men – to a greater than average degree. It may be supposed there are evolutionary causes for the latter (*Baumeister*, 2001). As an example of this phenomenon not strictly belonging to the field of economic science, a research study carried out in France revealed that 84% of French men declared themselves to be better than average lovers (*Dobelli*, 2013). Assuming a close to normal distribution on this question, in reality this figure would be around 50%

were it not for the existence of the overconfidence effect. Presumably we would obtain a similar result if we asked experts to position themselves and their level of knowledge within a distribution comprising their colleagues working in the same field of expertise.

Other factors also influence our thinking and thus the quality of our decisions. We will evaluate the same decision-making situation at two different times in different ways depending on our mood, the time of day and numerous other factors (e.g. our tiredness), which means that our value judgement is *incoherent* in time. Incoherent decisions are as typical of everyday situations as they are of decision-making situations supported by experts. A good example of this is a research study (*Hoffman–Slovic–Rorer*, 1968) in which experienced radiologists were asked to evaluate chest X-ray images and place them in the "normal" or "abnormal" categories at two different times, but without the participants in the experiment being aware that they were seeing the same images twice. In 20% of cases the assessments were contradictory. A similar degree of inconsistency was observed in a study in which 101 accountants were asked to evaluate the reliability of corporate internal audits (*Brown*, 1983).

WHEN CAN WE TRUST IN EXPERTS?

Having briefly reviewed the typical errors in thinking that may significantly distort experts' predictions and the formation of their judgements, the question arises: When and under what circumstances can we trust in an expert's opinion and intuition? To answer this question, we must take into account that expertise is really the accumulation of numerous relevant skills acquired over prolonged study. This is entirely obvious in the case of a professional chess player, for example. To acquire a high level of chess knowledge requires several hours of practice daily for a period of years. It is no different in any other specialised field. *Kahneman* states two basic conditions for the validity of expert value judgements (*Kahneman*, 2011):

- an environment that is sufficiently regular to be predictable;
- the opportunity to learn these regularities and acquire the appropriate skills through prolonged practice.

The first condition is obvious in the case of chess players, but engineers and physicians work in a similarly regular environment. By contrast, political scientists and securities analysts, who deal in long-term prognoses, essentially operate in a noisy, very irregular environment. Fulfilment of the second condition largely depends on the nature of feedback on the activity. *Kahneman* cites the example of medical specialists to throw light on this. "*Among medical specialties, anesthesiologists benefit from good feedback, because the effects of their actions are likely to be quickly evident. In contrast, radiologists obtain little information about the accuracy of the diagnoses they make and about the pathologies they fail to detect. Anesthesiologists are therefore in a better position to develop useful intuitive skills. If an anesthesiologist says, 'I have a feeling something is wrong,' everyone in the operating room should be prepared for an emergency.*" On this basis, it is also easy to understand why expert predictions are typically better on short time horizons than in the longer term. We simply do not have the power to practice longer-term forecasts often enough.

BRING ON THE STATISTICS!

Having reviewed some of the characteristics of human thinking and expert intuition that are definitive in decision-making situations, let us now examine what results can be obtained using mathematical and statistical tools.

The technological revolution and explosive development in the field of computing in recent decades means that we are now able to carry out several million operations within seconds with the help of personal computers. Previously, the application of complex mathematical and statistical algorithms (hereinafter: algorithms) tended in practice to be possible only within the framework of academic research, due to the significant time required to implement them and limited available computer capacity. Today, however, the situation has radically changed. A separate industry has emerged with the aim of making these algorithms accessible to users. One after another, various targeted software programs have become available which we can use to carry out analyses in a matter of minutes that might have once taken weeks. Thanks to open-source software, a significant portion of algorithms are now accessible to everyone.

The rapid pace of expansion and easy access to computing possibilities has naturally provided a further boost to underlying research – and within this the development of algorithms. New processes or algorithms regularly emerge from professional scientific workshops, of which the most viable are quickly adopted in practice.

In the area of applications, it was the corporate/business sphere that was the first to recognise the possibilities inherent in the new technology. In branches of industry characterised by production or services satisfying mass demand, the statistical analysis and modelling of business processes is a matter of course. In situations requiring decisions affecting masses, large modern companies were quick to realise that, with the help of predictive decision-making rules or formulae generated as a result of algorithms, they could reach decisions quickly and – perhaps even more importantly – cheaply.

In situations when the complexity and "size" of the problem under discussion is significant, algorithms can – by virtue of the very complexity of the decision-making situation – uncover correlations which the human mind can no longer grasp. This is particularly true when, for example, numerous factors impact the evolution of the quantity to be predicted, but individually these factors have comparatively little bearing on the end result. Take, for example, a credit institution with half a million customers and a correspondingly huge amount of descriptive data, where they may seek to answer the question: What determined whether or not someone repaid a loan, and how? Or, is it possible, in the case of a new customer, to estimate the probability of non-payment, and to decide who receives a loan on this basis?

With the help of appropriate statistical models, it is generally possible to reveal all the factors that exercise an influence on customers' inclination to pay, and – applying suitable algorithms – to generate a formula which can be used to estimate the likelihood of non-payment with respect to a specific customer. Based on the correlations thus revealed, it is possible to establish rules for the making of business decisions which, by means of an objective function, result in an optimal process, and which are automatically carried out during application.[1]

Given the complexity of the above decision-making situation, it is perhaps not surprising that in such areas human (expert) decisions/estimates perform worse than the algorithm-based rule. What is surprising, however, is that in many cases the same can also be observed for "small samples," where the problem can be described with considerably fewer data. In his 1954 work entitled *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, *Paul Meehl* summarised the findings of 20 research studies which examined whether the clinical predictions of trained experts based on subjective value judgements were more accurate than statistical forecasting that can be derived from a given rule. The results of the research showed that statistical predictions were typically far more reliable than the estimates of experts. Still more surprising are the findings on the topic by *Daniel Kahneman*, winner of the 2002 Nobel Prize in Economics, who writes in his book *Thinking, Fast and Slow*, published in 2011: "*The number of studies reporting comparisons of clinical and statistical predictions has increased*

1 With respect to the above two questions, a good example is the application of automatic credit assessment or so-called scoring systems used by major banks. Based on the data provided, these systems estimate the probability of the loan applicant's non-payment within seconds, entirely automatically and completely excluding human factors, and on this basis make a recommendation for the signing of the loan contract or – should there be a high probability of non-payment – rejection of the application.

to roughly 200, but the score in the contest between algorithms and humans has not changed. About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgment. No exception has been convincingly documented" (p. 181.). Another unexpected result of the research is that the superiority of statistical predictions can be seen even in the case of rules arising as a simple linear combination of explanatory factors.

The superiority of statistical predictions can also be found in areas where we would least expect it. The statistical prediction of Princeton economist and wine enthusiast *Orley Ashenfelter* (2007) regarding the development of prices of Bordeaux wines surpassed the prognoses of the most acclaimed wine experts. *Ashenfelter* employed data available in the year of the wines' production in making his forecast. Interestingly, his final predictive regression model contained three weather factors: the average temperature in the summer cultivation period, the quantity of rainfall at harvest time, and the total amount of rainfall over the preceding winter. Using only these three explanatory variables, *Ashenfelter* was able to explain the variance of future movements of Bordeaux wine prices at R^2=82%.

The extent to which *Ashenfelter*'s approach is not an isolated phenomenon is demonstrated by a number of research studies carried out in recent decades, mentioned here only in list form, which compare statistical and expert predictions on a given prognostic issue.

- *Howard* and *Dawes* (1976) applied the best and simplest possible method to predict the expected stability of marriages by the following formula: P = [frequency of lovemaking] minus [frequency of quarrels]. The reliability of the formula was later verified by Edwards (1977) and *Thornton* (1979).
- *Wittman* (1941) applied a statistical model to predict the success of electroshock therapy, the predictive capacity of which surpassed the prognoses of psychiatrists applying the procedure.
- Using statistical predictive methods, *Carroll* (1988) predicted repeat offences by convicted prisoners more successfully than experienced expert criminologists.
- In the case of certain students, statistical models drawn up at the time of admission provided better forecasts of their likely progress in university studies – be it medical school (*DeVaul*, 1957) or legal training (*Swets–Dawes–Monahan*, 2000) – than the teachers conducting the admissions process.
- In the area of lending by banks, the preference for statistical models over credit experts in predicting default on repayment was confirmed in studies by both *Stillwell* (1983) and *Libby* (1976).

- Several studies have indicated that the occurrence of early infant mortality (cot death) could be better predicted using statistical models than by experts (*Lowry,* 1975; *Carpenter*, 1977; *Golding,* 1985).
- The use of statistical models has proven better at predicting the likelihood of a given individual committing a crime in future than psychiatrists with expertise on the topic (*Faust* and *Ziskin*, 1988).

Beyond their typically better performance, models that serve as the basis for statistical predictions are also suitable for uncovering the explanatory factors of a prognosis, as well as the correlations between these factors and their relationship with the quantity to be predicted. In this way, they not only serve as a well-functioning "crystal ball" in forecasting, but can also help us to acquire knowledge and understanding of the internal mechanism and network of correlations of the phenomenon to be modelled. To put it simply, we could also say that, with an adequate database, a good statistical model is able to "learn" within a few minutes everything which an expert in the field might have needed several decades to learn.

THE VALIDITY OF STATISTICAL PREDICTIONS

The foregoing might perhaps lead us to conclude that, in most decision-making and predictive situations, we can entirely dispense with human wisdom or expert knowledge. "Luckily," even statistics-based algorithmic decision-making and predictive systems are not without certain disadvantages. One is the situation described by *Paul Meehl* as the "broken leg" phenomenon. In his thought experiment, Meehl supposed that we have a statistical algorithm which, based on earlier experiences, is able to predict with a great degree of certainty that a certain professor will go to the movies on Wednesday evening. The algorithm works superbly until one Tuesday the professor unexpectedly breaks his leg, and thus cannot go to the movies on Wednesday. This shows that algorithms perform poorly in situations when a rare, previously unobserved event of low probability occurs which exercises a significant effect on the eventual outcome.[2] The other typical problem is related to the data structure that represents the starting-point of algorithm development. Predictive models are only able to provide optimal results within the world represented by the available data. If, for some reason, essential variables that strongly influence the development of the quantity to be predicted are omitted from the database, then the quality of the prediction will deteriorate. An example of a situation such as this is when a bank, in conforming to data protection

2 In short-term economic forecasts, atypical events such as this include the trend reversal or the crisis situation.

regulations, is unable to collect certain types of data from clients (e.g. on religion, race, etc). Another important aspect in practical application is the question of the reliability and quality of available data, which is primarily determined by data recording and management processes. The quality and reliability of the data used is definitive from the point of view of a model's predictive capacity. If we use unrealistic data structures that are tainted with erroneous data during our statistical modelling, then the results, too, will carry limited validity. This phenomenon is conveyed – perhaps somewhat exaggeratedly – by the so-called GIGO principle (Garbage In, Garbage Out).[3]

One barrier to the application and spread of statistical models, which is natural and boils down to human psychology, is the antipathy and disapproval often shown towards them. A good example of this is the reception given to the aforementioned formula *Ashenfelter* developed to forecast the prices of Bordeaux wines. The reactions of French wine-lovers, according to *The New York Times*, ranged "between violent and hysterical." *Kahneman* notes that prejudice against algorithms is further increasing, given that algorithm-based decisions have significant consequences. For most people, after all, it does matter whether the source of decision-making errors that affect real-life situations is an algorithm or an expert. In the case of medical malpractice, for example, it is more upsetting if the death of a patient occurs due to the application of a formula than if a doctor's bad decision has led to the unfortunate event.

ALGORITHMS OR EXPERTS?

In conclusion, let us attempt to review whether we should resort to statistical tools in a given decision-making situation, or if we should rely on the advice of experts instead.

In a noisy, loosely structured predictive environment – such as the world of securities prices, for example – the lesson of research is that neither mathematical/statistical algorithms nor expert predictions perform adequately, irrespective of the time horizon. In this case, the biggest problem is the lack of regularities and patterns in the observed system, which consequently neither an algorithm nor an expert is able to master.

In more or less regular environments (such as, for example, healthcare applications), with adequate data as a basis and using the appropriate algorithms, we can hopefully expect better forecasts on both short and longer predictive time hori-

3 Certain statistical models are suitably robust with respect to the distribution of the employed variables.

zons alike than if we rely upon expert estimates. In the longer term, the problem for experts is that there is no opportunity for an adequate amount of "practice" to deepen their knowledge. In the case of algorithms, it is generally the accessibility of data, the quality of the data structure and limits on the backtesting (verification) of models which may present obstacles.

In a regular, well-structured environment, the primacy of algorithmic forecasting is indisputable based on research, provided there is a suitable quantity and quality of available data relevant from the point of view of the problem concerned. In this case, algorithmic predictions are substantially better than expert estimates. An additional major advantage is that, rendered automatic in mass decision-making, they can be applied cheaply and quickly. Empirical confirmation of this can be best observed in sectors that provide mass services (banking, insurance, telecoms, etc), in the case of rapidly proliferating automated credit scoring, CRM, CHURN and other statistically based decision-making systems. Beyond their typically better predictive performance, models that serve as the basis for statistical predictions are also suitable for uncovering the explanatory factors of a prognosis, as well as the correlations between these factors and their relationship with the quantity to be predicted. In this way, they are useful not only for prognostic purposes, but can also help us to acquire knowledge, new skills and understanding of the internal mechanism and network of correlations of the phenomenon to be modelled.

REFERENCES

Ashenfelter, Orley (2007): *Predicting the quality and prices of Bordeaux wines*, AAWE working paper, No. 4.

Baumeister, Roy F. (2001*):Is there anything good about men? How culturesflourish by exploiting men.* Oxford University Press 2001, 211.

Brown, Paul R. (1983): Independent auditor judgment inthe evaluation of internal audit functions. *Journal of Accounting Research* 21: 444–455.

Carpenter, R. G. – Gardner, A. – McWeeny, M. – Emery, J. L. (1977): Multistage scoring system for identifying infants at risk of unexpected death. *Arch. Dis. Childh.,* 53: 606–612.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J. – Alibrio, J. J. (1982): Evaluation, diagnosis, and prediction in parole decision-making. *Law and Society Review*, 17: p

DeVaul, R. A. – Jervey, F. – Chappell, J. A. – Carver, – Short, B. – O'Keefe, S. (1957): Medical school performance of initially rejected students. J*ournal of the American Medical Association* 257: 47–51.

Dobelli, Rolf (2011): *Die Kunst des klaren Denkens: 52 Denkfehler, die Sie besser anderen* überlassen. München: Carl Hanser Verlag GmbH & Co.

Golding, Jean – Limerick, Sylvia – MacFarlane, Aidan (1985): *Sudden infant death*. Somerset: Open Books.

Hoffman, Paul J. – Slovic, Paul – Rorer, Leonard G. (1968): An analysis-of-variance model for the assessment of configural cue utilization in clinical judgement. *Psychological Bulletin* 69: 338–339.

Howard, J. W. – Dawes, R. M. (1976): Linear prediction of marital happiness. *Personality and Social Psychology Bulletin* 2: 478–480.

Kahneman, Daniel (2013): *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux. file:///D:/gp/Thinking,%20Fast%20and%20Slow%20-%20Daniel%20Kahneman.pdf.

Libby, R. (1976): Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance* 16: 1–12.

Lowry, C. (1975): *The identification of infants at high risk of early death.* Med. Stats. Report, London School of Hygiene and Tropical Medicine.

Meehl, (1954): *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, MN: University of Minnesota Press.

Slovic, (2000): *The perception of risk*. London: Earthscan.

Swets, J. A. – Dawes, R. M. – Monahan, J. (2000): Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest* 1: 1–26.

Stillwell, W.G. – Barron, F. H. – Edwards, W. (1983): Evaluating credit applications: A validation of multiattribute utility weight elicitation techniques. *Organizational Behavior and Human Performance* 32: 87–108.

Tetlock, Philip E. (2006): How accurate are your pet pundits? Project Syndicate/Institute for human sciences.

Wittman, P. (1941): A Scale for Measuring Prognosis in SchizophrenicPatients. Elgin Papers 4: 20–33.