On the role of psychological heuristics in operational research; and a demonstration in military stability operations

Niklas Kellera,b, Konstantinos V. Katsikopoulosb,*

a Charité University Hospital, Department of Anesthesiology, Charitéplatz 1, 10117 Berlin, Germany
b Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition (ABC), Lentzeallee 94, 14195 Berlin, Germany

ARTICLE INFO

Article history:
Received 11 November 2014
Accepted 8 July 2015
Available online xxx

Keywords:
Behavioural OR
Bounded rationality
Heuristics
Decision analysis
Forecasting

ABSTRACT

Psychological heuristics are formal models for making decisions that (i) rely on core psychological capacities (e.g., recognizing patterns or recalling information from memory), (ii) do not necessarily use all available information, and process the information they use by simple computations (e.g., ordinal comparisons or un-weighted sums), and (iii) are easy to understand, apply and explain. The contribution of this article is fourfold: First, the conceptual foundation of the psychological heuristics research program is provided, along with a discussion of its relationship to soft and hard OR. Second, empirical evidence and theoretical analyses are presented on the conditions under which psychological heuristics perform on par with or even better than more complex standard models in decision problems such as multi-attribute choice, classification, and forecasting, and in domains as varied as health, economics and management. Third, we demonstrate the application of the psychological heuristics approach to the problem of reducing civilian casualties in military stability operations. Finally, we discuss the role that psychological heuristics can play in OR theory and practice.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

1. Introduction

If operational research (OR) methods are developed by people and are to be implemented by other people who work with yet others, how can effective operational research be anything else than behavioural? It cannot. Recently, an article explicitly calling for more behavioural research in OR gained traction (Hamalainen, Luoma, & Saarinen, 2013; see also Bearden & Rapoport, 2005). Furthermore, the British OR Society established a special interest group in behavioural OR and a plenary panel on behavioural OR was organized in the Society’s annual conference (Montibeller et al., 2014).

The basic idea is to import methods and knowledge from the behavioural sciences, notably the psychology of judgment and decision making and behavioural economics, to OR. An important discovery of these sciences is that people exhibit bounded rationality. Bounded rationality refers to situations in which there is not enough time or other resources, such as computational capacity, to obtain all information and find an optimal solution, but nevertheless a good feasible solution must be identified. In other words, bounded rationality is the kind of rationality that most of us, laypeople and experts, realistically need to exhibit in our life and work.

Herbert Simon (1955, 1998), one of the great 20th century polymaths—who sometimes also wore the hat of an operational researcher— is credited as the father of the idea of bounded rationality but refrained from giving a precise definition. Thus, there are multiple views of bounded rationality, as is often noted (Gigerenzer & Selten, 2001; Katsikopoulos, 2014; Lee, 2011; Rubinstein, 1998).

This article critically discusses one of these views of bounded rationality, a view that we consider as particularly relevant to OR. This view has a strong behavioural component: It consists of prescriptive models of decision making which have also been used to describe people’s actual behaviour. These models include the few pieces of information that people use and specify the simple ways in which they process this information. The models go under labels such as “fast-and-frugal heuristics” (Gigerenzer, Todd, & the ABC research group, 1999), “simple models” (Hogarth & Karelaia, 2005), “psychological heuristics” (Katsikopoulos, 2011) and “simple rules” (Eisenhardt & Sull, 2001). In this article, we will use the term psychological heuristics.

The contribution of the article is fourfold: The conceptual foundation of the psychological heuristics research program, along with a discussion of its relationship to soft and hard OR, is provided in Section 2. In Section 3, empirical evidence and theoretical analyses are presented on the conditions under which psychological heuristics perform on par with or even better than more complex standard models in decision problems such as multi-attribute choice,
classification and forecasting, and domains as varied as health, economics, and management. Sections 4–6 demonstrate the application of the psychological heuristics approach to the problem of reducing civilian casualties in military stability operations. Section 7 summarizes the main messages of the article and discusses the role that psychological heuristics can play in OR theory and practice.

2. The conceptual foundation of psychological heuristics and its relationship to OR

There are three interpretations of heuristics which are relevant to this article. First, in hard OR and computer science, heuristics refer to computationally simple models which allow one to "...quickly [find] good, feasible solutions" (Hillier & Lieberman, 2001, p. 624). These heuristics can be developed from formal principles or be empirically driven. The other two interpretations of heuristics come from psychology. Kahneman, Slovic, and Tversky (1982) focused on the experimental study of psychological processes that "in general...are quite useful, but sometimes lead to severe and systematic errors" (Tversky & Kahneman, 1974, p. 1124) and proposed informal models (i.e., models that do not make precise quantitative statements) of heuristics. Gigerenzer et al. (1999) developed and tested formal models of heuristics that, they argued, "...when compared to standard benchmark strategies,... can be faster, more frugal, and more accurate at the same time" (Gigerenzer & Todd, 1999, p. 22).

Katsikopoulos (2011) proposed a definition of psychological heuristics which is a hybrid of these three interpretations. As in Tversky and Kahneman (1974) and Gigerenzer et al. (1999), this definition focuses on heuristics that not only are computational shortcuts but also have a psychological basis; and as in Hillier and Lieberman (2001) and Gigerenzer et al. (1999), these heuristics are formalized. Psychological heuristics are formal models for making decisions that

(i) rely on core psychological capacities (e.g., recognizing patterns or recalling information from memory).
(ii) do not necessarily use all available information, and process the information they use by simple computations (e.g., ordinal comparisons or un-weighted sums), and
(iii) are easy to understand, apply and explain.

As they are stated above, requirements (i), (ii) and (iii) are partly underspecified but the following discussion should clarify their meaning. Consider the problem of choosing one out of many apartments to rent based on attributes such as price, duration of contract and distance from the centre of town. The standard approach of hard OR, decision analysis (Keeney & Raiffa, 1976), includes eliciting attribute weights, single attribute functions and interactions among attributes. Such different pieces of information are then integrated by using additive or multi-linear functions. On the other hand, a psychological heuristic for solving the problem could be to decide based on one attribute—say, price—or order attributes by subjective importance and decide based on the first attribute in the order which sufficiently discriminates among the alternatives (Hogarth & Karelaia, 2005; Tversky, 1972).

For example, price could be ranked first and contract duration second, and prices could differ only by 50 Euros per month while contract durations could differ by a year, in which case the apartment with the longest contract would be chosen (assuming that you prefer longer to shorter contracts). In a review of 45 studies, Ford, Schmitt, Schechtman, Hults, and Doherty (1989) found that people very often use such heuristics for choosing items as diverse as apartments, microwaves and birth control methods.

As a second example, consider the problem of forecasting which one of two companies will have higher stock value in five years from now. Assuming that you recognize only one of the two companies, a psychological heuristic for making such decisions is to pick the recognized company (Goldstein & Gigerenzer, 2009). This is in stark contrast with doing the computations of mean-variance portfolio optimization (Markowitz, 1952).

As said, psychological heuristics differ from the heuristics of the "heuristics-and-biases" research program (Kahneman et al., 1982) in that they are models which make precise quantitative statements. An important distinction is that psychological heuristics are not only descriptive of actual behaviour but are also argued to be prescriptive under some conditions (see Section 3), whereas Kahneman et al. (1982) heuristics are meant to be descriptive and prescription enters through counteracting biases (Edwards & Fasolo, 2001). For further discussion, see Kelman (2011) and Katsikopoulos and Gigerenzer (2013).

Formal modelling also differentiates psychological heuristics from the "naturalistic decision making" research program (Zsambok & Klein, 1997). There exist, however, a number of points of synergy of these two approaches (Keller, Cokely, Katsikopoulos, & Wegwarth, 2010). For example, heuristics are not automatically assumed to be second best to a theoretically optimal solution.

Psychological heuristics target some of the same type of problems to which hard OR has been applied. In these problems, there is a clear objective (e.g., choose the company with the higher stock value five years from now) and the success of a method may be evaluated by using standards such as agreement with the ground truth (e.g., company stock values). Like hard OR methods, heuristics are formal models, and thus differ from a restatement or reuse of managerial intuition.

On the other hand, psychological heuristics differ from heuristics of hard OR in that they not only are computational shortcuts but also have an identifiable psychological basis. This psychological basis can be due to expertise (Klein, 1999). For example, some experienced managers are aware of the fact that customers who have not bought anything from an apparel company in the last nine months are very unlikely to buy something again in the future, and use this single attribute to make more accurate decisions about targeted advertising than using a standard forecasting model (Wuebben & von Wangenheim, 2008). Psychological heuristics can also be grounded in processes available to laypeople. For example, a human child can recognize faces better than currently available software (with the possible exception of new anti-terrorist technologies).

Some heuristics of hard OR may, at the formal level, look like the heuristics a person would spontaneously use as in solving the traveling salesman problem by going to the closest unvisited town. But the process of arriving at the heuristics is different. Unlike hard OR models, psychological heuristics are not derived by solving or approximating for the optimal solution of a model. Rather, psychological heuristics are based on observation and analysis of human behaviour, and in particular of how people make good decisions with little data.

Psychological heuristics have a nuanced relationship with methods of soft OR (Rosenhead & Mingers, 2001). The main point is that psychological heuristics and soft OR methods are not used for the same type of problems. Unlike soft OR, the heuristics discussed in this article do not apply to wicked problems (Churchman, 1967) with unclear objectives or multiple disagreeing stakeholders. Success of soft OR methods may mean that communication among stakeholders was enhanced or consensus was achieved (Mingers, 2011), whereas the success of psychological heuristics may be measured quantitatively.

Furthermore, the characteristics of the offered solutions also differ. In soft OR, the solution may be a set of qualitative principles which allow objectives to be clarified and stakeholders to work together whereas psychological heuristics are formal models of actual effective behaviour. The process of deriving the solutions is, in both cases, based on the observation and analysis of people's behaviour, but in soft OR the focus is typically on counteracting biases whereas psychological heuristics focus on how good decisions are made.
Table 1
A brief summary of some conceptual connections among soft OR, psychological heuristics and hard OR, as discussed in Section 2. It can be argued that psychological heuristics lie between soft and hard OR.

<table>
<thead>
<tr>
<th>Target problems</th>
<th>Soft OR</th>
<th>Psychological heuristics</th>
<th>Hard OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclear objectives, multiple disagreeing stakeholders, success may mean enhancing communication or achieving consensus</td>
<td>Clear objectives, individual decision makers, success may be measured by agreement with ground truth</td>
<td>Clear objectives, success may be measured by agreement with ground truth</td>
<td></td>
</tr>
<tr>
<td>Process of deriving solutions</td>
<td>Observe and analyse people’s purposeful behaviour, aiming at counteracting biases</td>
<td>Observe and analyse people’s behaviour, when they made good decisions with little data</td>
<td>Solve or approximate the optimal solution of a model, not using knowledge of people’s behaviour</td>
</tr>
<tr>
<td>Characteristics of solutions</td>
<td>Qualitative principles which allow objectives to be clarified and stakeholders to work together</td>
<td>Models of people’s effective behaviour, formalized so that they conform to (i), (ii) and (iii) in the definition</td>
<td>Models, not descriptive of people’s behaviour, meant to improve on unaided intuition</td>
</tr>
</tbody>
</table>

It is noteworthy that there is a crucial point of convergence of psychological heuristics and soft OR. Both approaches acknowledge the possibility that high-quality data, say on utilities or probabilities, may be missing, and tailor their methods accordingly. This will be especially important in the military stability operations problem we consider in Sections 4–6.

The above points are summarized in Table 1. In sum, it can be argued that psychological heuristics lie somewhere between hard and soft OR, and in this sense could be used to bridge the gap between them.

3. Models and performance of psychological heuristics

3.1. Models of psychological heuristics

A main family of psychological heuristics is that of lexicographic models (Fishburn, 1974). An example of a lexicographic model is the apartment-renting heuristic of Section 2: Order apartment attributes by subjective importance and make a decision based on the first attribute in the order which sufficiently discriminates among apartments. Note that a heuristic which only uses one attribute (e.g., if you recognize the name of a company, invest on it), is also a lexicographic model.

Lexicographic models have been applied to problems of multi-attribute choice, classification and forecasting. In multi-attribute choice, the objective is to choose one out of many alternatives, which offers the maximum true multi-attribute utility to the decision maker, as for example overall satisfaction from renting an apartment.

In classification, the objective is to classify an object to one out of many possible categories, again based on its attribute values. For example, a classification problem is to decide if a patient with some known symptoms, such as intense chest pain, is at a high risk of a heart attack and needs to be in the emergency room or should communicate to patients (Elwyn, Edwards, Eccles, & Rovner, 2001).

Furthermore, financial practitioners have also started using fast and frugal trees. For example, economists from the Bank of England have developed a fast and frugal tree for forecasting whether a bank is at risk of bankruptcy or not (Aikman et al., 2014). When compared with the standard of financial economics, logistic regression, the tree was not outperformed by any of the 20 versions of logistic regression—which used the same economic indicators as the tree—while being much easier to understand and use.

In fast and frugal trees and other lexicographic models, attributes may be ordered by a measure of the statistical correlation between each attribute of the object and the utility or category of the object (Luan, Schoolder, & Gigerenzer, 2011). This means that data on attributes and utilities or categories of objects is required. The set containing this data is called the training set. It has been found that when people are given a training set of adequate size, and long enough time to learn from it, they can order attributes by their correlation with the specified criterion (Broder & Newell, 2008).

It has been argued that fast and frugal trees make the medical decision process more transparent and easier to understand and communicate to patients (Elwyn, Edwards, Eccles, & Rovner, 2001).

In fast and frugal trees and other lexicographic models, attributes may be ordered by a measure of the statistical correlation between each attribute of the object and the utility or category of the object (Luan, Schoolder, & Gigerenzer, 2011). This means that data on attributes and utilities or categories of objects is required. The set containing this data is called the training set. It has been found that when people are given a training set of adequate size, and long enough time to learn from it, they can order attributes by their correlation with the specified criterion (Broder & Newell, 2008). It is important to note, however, that fast and frugal trees do not necessarily require statistical data to be built. Another way of building a fast and frugal tree is to...
combine expert knowledge with task analysis (Vicente, 1999). In fact, this is the way by which we built a fast and frugal tree for reducing civilian casualties in military stability operations (see Section 6).

In forecasting problems, the ground truth is not known now but will be available in the future (e.g., company stock values in five years). Forecasting does not necessarily refer to making point estimates (e.g., predicting the stock value of a company in five years). Rather, forecasting could also mean making multi-attribute choices (e.g., which one of two companies will have a higher stock value in five years?) or classifications (e.g., will this company be bankrupt within five years?) into the future. More generally, any data point which is not known now can be viewed as being in the future.

There are many other models of psychological heuristics beyond lexicographic ones (Gigerenzer, Hertwig, & Pachur, 2011). Another main family of heuristics is that of tallying (or unit-weights) models (Dawes & Corrigan, 1974). Tallying models are linear models for multi-attribute choice, classification and forecasting, where the weights of all attributes are set to 1. Surprisingly, it has been found in applications in psychometrics and personnel selection, that tallying could sometimes forecast better than linear regression with unconstrained attribute weights (Bobko, Roth, & Buster, 2007). Tallying also could not be outperformed by 13 versions of Markowitz’s mean-variance optimization model in allocating wealth across assets in seven real financial portfolios (DeMiguel, Garlappi, & Uppal, 2007).

Finally, note that tallying and lexicographic models occupy the two extremes of a continuum. In tallying, each attribute can compensate for any other attribute, whereas in lexicographic models, the first discriminating attribute cannot be compensated by all other attributes put together. The few applications discussed in this section hint that psychological heuristics may be able to compete with more complex models (Todd, 2007). But are these isolated incidents? The next sub-section provides a systematic review.

3.2. When do psychological heuristics perform well?

A fair amount of empirical evidence has accumulated on the comparison between psychological heuristics and more complex models typically used in statistics, computer science and hard OR such as regressions (linear, logistic and regularized), Bayesian networks (e.g., naïve Bayes), neural networks, classification and regression trees, and support vector machines; in the remainder of the article, we refer to these models as standard models. The focus of this review is on studies which used a range of decision problems to compare the performance of psychological heuristics and standard models.

First, Czerlinski, Gigerenzer, and Goldstein (1999) estimated the performance of linear regression, a lexicographic model and tallying on 20 datasets from the fields of biology, environmental science, demography, health, psychology and transportation. The average number of alternatives in each dataset was 67. In each dataset, all multi-attribute choices involving two out of a fixed half of the alternatives in the dataset were constructed and used as the set on which the models were trained (i.e., attribute weights in regression and the order of attributes in the lexicographic model were estimated). The other half of the choices was used to measure the predictive (forecasting) performance of the models, as the percentage of agreement with the ground truth. For example, say that a choice is to pick the one out of two American high schools with the higher dropout rate; choosing the high school which indeed has the higher rate agrees with the ground truth. The sampling of the training set was repeated thousands of times to average out random variation.

The main finding of the Czerlinski et al. (1999) study is that, averaged across the 20 datasets, the predictive accuracy of lexicographic model equals that of linear regression, 76 percent, whereas tallying scores 69 percent. This result has been replicated and also extended to include other lexicographic and standard models (e.g., naïve Bayes) and very small training sets, which may be more realistic for decision making under changing conditions. Katsikopoulos, Schooler, and Hertwig (2010) tested training sets with sizes ranging from 3 percent to 15 percent of the datasets on average. Their main findings are that tallying is the best-performing method for the smallest training set and that the best lexicographic model is the best method for the other training sets, outperforming naïve Bayes by more than 5 percent on the average (naïve Bayes scored 73 percent on the 50 percent training set on the average).

Second, Martignon et al. (2008) estimated the performance of classification and regression trees, logistic regression and two versions of fast and frugal trees, on 30 datasets from the UC Irvine Machine Learning Repository. All datasets referred to classifications, and 11 referred to medical classifications. The procedure of Czerlinski et al. (1999) was used, with the size of the training set being 15 percent, 50 percent or 85 percent of each dataset. The main finding is again a strong effect of training set size: In prediction (forecasting), the best standard model outperforms the best psychological heuristic for the largest training set by 4 percent (82 percent vs. 78 percent), but the difference shrinks to 1 percent (76 percent vs. 75 percent) for the smallest training set. Brighton (2006) obtained similar results, using eight of the problems of Czerlinski et al. (1999), and also testing Quinlan’s (1990) C4.5 decision trees.

Finally, in another study looking at a broad range of problems, Şimşek (2013) extended the design of Czerlinski et al. (1999), by using their 20 datasets plus 31 more, spanning the fields of biology, business, computer science, ecology, economics, education, engineering, environmental science, medicine, political science, psychology, sociology, sports and transportation. She also used a state-of-the-art version of linear regression with elastic net regularization plus a number of lexicographic models. The size of the training set was equal to the size of the whole dataset minus one. The main finding is that linear regression scores 79 percent and the best lexicographic model 78 percent. This study also reported results on each dataset; note that the absolute and relative performance of methods varies greatly across datasets (e.g., the performance of the best lexicographic model varies from 40 percent to 92 percent).

A tentative summary of the empirical evidence is the following:

(1) Overall, there are no large performance differences between psychological heuristics and standard models. (2) Both psychological heuristics and standard models can outperform each other for particular datasets and (3) psychological heuristics fare especially well in problems of prediction/forecasting.

(1) can be explained by the flat maximum effect (Lovie & Lovie, 1986). Informally, this effect says that the attribute weights used in a linear model do not change much the overall deviation between true and forecasted values. This is relevant because many of the psychological heuristics and standard models tested can indeed be viewed as linear models. For linear regression and tallying, this is so by definition; for naïve Bayes and lexicographic models, see Katsikopoulos (2011).

(2) and (3) can be explained by the framework of the statistics of prediction, and in particular by the bias-variance decomposition of prediction error (Geman, Bienenstock, & Doursat, 1992). Informally, this decomposition says that the controllable prediction error of a method is the sum of two terms—the bias term which measures how well can the method on the average agree with the ground truth, and the variance term which measures the variation around this average.

Gigerenzer and Brighton (2009) conjectured that psychological heuristics tend to have higher bias but lower variance than standard models. The claim of lower variance can be intuited in the case of small training set sizes: The smaller the training set, the more likely it is that sampling error and natural variations in the instances included in the training set will lead to variation in the parameter estimates of a given method. This variation can be expected to influence the more
heavily parameterized standard models to a greater degree than the simpler psychological heuristics.

Recently it was discovered that, under some conditions, psychological heuristics may also have bias competitive to that of standard models. These conditions express structures which guarantee the optimal performance of lexicographic models (when true utility is an additive or multi-linear function of the attributes): The presence of a non-compensatory attribute or of a, simply or cumulatively, dominant alternative (Katsikopoulos, 2011). A non-compensatory attribute is one which has a much higher correlation with the ground truth than all other attributes in a sense put together (for formal details, see Katsikopoulos & Martignon, 2006; Martignon & Hoffrage, 2002). A dominant alternative is one which has superior or equally good values on all attributes, compared to all other alternatives in the normal or cumulative attribute profile (for formal details, see Baucells, Carrasco, & Hogarth, 2008).

One may expect that these expectations are rare in the real world. This is not the case. Şimşek (2013) found that non-compensatory attributes exist in 93 percent of her binary datasets (i.e., attributes had values of 1 or 0) and 83 percent of the numeric datasets; and that cumulatively dominant alternatives exist in 87 percent and 58 percent of the binary and numeric datasets respectively. For further analyses, see Baucells et al. (2008) and Katsikopoulos, Egozcue, and Garcia (2014).

Table 2 below summarizes informally the empirical findings of the studies using broad ranges of decision problems, as well as the theoretical explanations of these findings.

The next three sections of the article demonstrate how to develop a psychological heuristic and show empirically the benefits of doing so. We consider the important problem of reducing civilian casualties in military stability operations.

4. The problem of reducing civilian casualties in military stability operations

Modern stability operations (e.g., in Afghanistan or Kosovo) have a very different character from that of classic interstate warfare. Classic interstate warfare aims at the elimination or subjugation of existing political and military institutions. Stability operations aim at creating a stable and secure environment that allows for the creation of such institutions in the first place. Beyond being morally imperative, the minimisation of civilian casualties is therefore a central strategic concern in stability operations. Any loss of life on the part of the civilian population not only increases local resentment, swelling the ranks of resistance forces, but is also a direct failure to fulfil the political mandate of the operation.

Consequently, there has been increasing pressure to minimise civilian casualties, for example in Afghanistan, where a number of strategic directives have been issued to that effect (McChrystal, 2009; Petraeus, 2009). These have been successful in some types of operations in which the decision to use force has to be approved by higher levels of command. In close air support, for example, the civilian death toll has decreased from 359 in 2009, to 171 in 2010 and 187 in 2011 (United Nations Assistance Mission to Afghanistan [UNAMA], 2009, 2010, 2011). Yet, in situations in which individual soldiers have to assess potential threats and judge how to react appropriately under time pressure, no such reductions have been forthcoming. In the three years following the introduction of the new tactical directives, civilian deaths in such situations have not decreased as 36, 45 and 38 civilians were killed in these years (UNAMA, 2009, 2010, 2011). The goal of our work was to construct a usable and effective decision aid for soldiers.

In order to frame the problem, we employed methods of cognitive field research, specifically goal-directed task analysis (Hoffman, 2005), which consisted of reviews of the academic and practitioner literature, observations of military training and interviews with armed forces instructors and other experienced personnel. Initially, the idea was to only look at checkpoints. However, the analysis revealed that the scope could be expanded to encompass all types of force protection incidents including military presences such as checkpoints, convoys and ground patrols. We decided to only address the situation of approaching vehicles and not persons approaching a military presence on foot. The reason was that such situations differ greatly in the attributes that can be used for classification as well as the dynamics of the situation (e.g., the speed of approach or the potential size of the blast radius of vehicle borne suicide attackers). After framing the problem with the help of the goal-directed task analysis, we looked at the common characteristics across all types of force protection incidents.

Common across all force protection situations is the legal framework that governs the use of force. NATO soldiers in stability operations are typically under at least three sets of legal constraints; of their country of origin, international law and of the host nation. These legal constraints are translated into operational guidelines called rules of engagement, which govern the use of force. Soldiers are given laminated pocket cards with the most operationally relevant rules of engagement and which are for all practical purposes binding orders for soldiers in the field. For example, one rule is that soldiers may use lethal force only for extended self-defence or if a hostile has been positively identified.

Yet, these pocket cards do not address some important concerns. The soldiers do not know if they are actually under threat and if so, what the nature and level of that threat is and how to best respond to it. To help soldiers deal with such uncertainties, NATO has developed a further set of guidelines called the Escalation of Force chain (EoF chain). An EoF chain is a sequence of actions, increasing in lethality, for guiding the application of force until a possible threat

---

Please cite this article as: N. Keller, K.V. Katsikopoulos, On the role of psychological heuristics in operational research; and a demonstration in military stability operations, European Journal of Operational Research (2015), http://dx.doi.org/10.1016/j.ejor.2015.07.023
is deterred or eliminated. The official NATO EoF guidelines for the application of force to oncoming traffic are shown in Fig. 2.

These guidelines are the same irrespective of the particular situation the soldier is in (convoy, checkpoint or ground patrol). To signal military presence, soldiers always employ level EoF 1. In practice, this may include auditory (e.g., shouting, bullhorn, loudspeakers or car horn) or visual (e.g., stop signs, vehicle lights or waving) signals. EoF 2 also includes the use of nonlethal means such as signal flares or dazzling lasers. In EoF 3, a warning shot may also be given off into the air. A disabling shot may be placed into the tires in EoF 4. In the final level EoF 5, shots are used to make sure that the threat is eliminated. Levels may be skipped, for example if there is time pressure, or they may be repeated.

The problem is that soldiers, across all types of force protection situations, have practically no guidance on how to use the EoF chain. They are simply told: “If time and circumstances permit…” They have no instructions on whether to escalate to a higher EoF level, when to do so or up to which level to escalate. They are left alone to identify the relevant attributes and to combine them in order to decide when and how to escalate force.

5. A descriptive model of soldier decision-making in force protection problems

5.1. Data

In order to determine the attributes soldiers use, we investigated soldiers' decisions and actions in 1060 reports of incidents involving motor vehicles approaching a military presence between January 2004 and December 2009 in Afghanistan across all six ISAF regional commands (Wikileaks, 2010). The goal-directed task analysis was used to construct a coding schema for these reports and five categories of importance to the problem were identified: (i) the EoF actions employed (beyond EoF 1), (ii) type and number of casualties (injured or killed), (iii) situational information (type of vehicle; moving or stationary), (iv) frequent attributes (speed, proximity or number of vehicle occupants, etc.) and (v) infrequent attributes (“lying low to the ground”, “weapons visible”, “coming under fire” or “intelligence information”, etc.) An example report is given below where the information that was coded is italicized. Some information in this example has been removed as it was specific enough to potentially allow identification of the particular unit and persons involved in the incident. The removal of this information does not impact our analysis in any way.

At XXXX TF (task force) Kandahar reports XXXXXXXXXXXXXX section injured 2xLNV's (local nationals) at XXXXXXX in XXXXXXX district of XXXXX province. A white Toyota corolla [sic] with 2xLNV's in it was approaching too fast, FF [friendly forces] told vehicle to stop a number of times, vehicle failed to stop, FF engaged vehicle wounding 1xLNV in car and 2ndLNV (child) ran into mosquito wounded.

UPDATE: FF engaged the vehicle and took cover because of recent J2 (intelligence) indications of a possible Toyota corolla SVBIED (suicide vehicle-borne improvised explosive device) operating in their current AO (area of operations). XX MP (military police) are on scene and collecting details. ISAF tracking XXXXXX.

Note that the reports were written only after EoF levels 2 or higher were employed and therefore the dataset does not include those non-hostile incidents that were classified as soon as the vehicle appeared or right after EoF 1 was employed. Furthermore, the reports did not discuss whether any hostile vehicles were deterred by the mere fact of a military presence. Finally, the reports did not typically discuss in sufficient detail the weapons platforms used even though this factor has been shown to have an important effect.

Originally, 1087 reports were available. Two raters coded the reports independently with an inter-rater agreement of 86.2 percent in the first round. The raters discussed all points of divergence and 27 cases for which no agreement could be found were excluded, leaving 1060 reports. In 1053 out of 1060 incidents, the vehicle occupants were not hostile and a total of 204 civilians were reported injured or killed. The other seven reports describe successful attacks on NATO forces by suicide attackers. Note that these seven reports discuss only the battle damage and do not mention any threat attributes or EoF levels.

5.2. An analysis of soldier behaviour

As said in Sections 2 and 3, human decision-making is often based on just one attribute (Ford et al., 1989). Thus, we started our analysis by looking for a single prominent attribute in the reports. Indeed, there is one attribute mentioned in all 1053 reports involving non-hostiles, the compliance attribute. An oncoming vehicle is complying if it is slowing down, stopping or moving out of the way while signalled by the soldier. In 1020 incidents where the traffic did not comply, the soldiers always increased the EoF level and only in four noncompliant incidents was the EoF chain aborted (in the other 29 noncompliant incidents, an EoF level was repeated). Conversely, as soon as the vehicle complied, the soldiers aborted force escalation every time.

The use of compliance as an attribute is sensible from the soldier's perspective. First, achieving compliance of an oncoming vehicle is in itself an operational objective. While staffing a checkpoint or defusing a roadside bomb, the soldier cannot afford to have traffic rush through the area as this may endanger the lives of the civilians or compromise military security. Second, compliance also increases the time to collision of the approaching vehicle and thus increases safety. Time to collision dictates the time the soldier has available to gather further information and assess threat as well as the time available to perform any necessary action. Finally, compliance may be perceived as having diagnostic value as it may be assumed that civilians will stop or slow down when detecting a military presence and wait for instructions whereas suicide attackers may rush toward the military presence before detonating.

What actions did the soldiers take based on the compliance attribute? Table 3 shows the number of times a particular level in the EoF chain was used, the number of times it was used as the highest level and the resulting number of civilian casualties.

Table 3 shows that EoF 4 and EoF 5 were used much less frequently than EoF 2 and EoF 3, which means that soldiers did not often decide to directly engage oncoming traffic. When soldiers did decide to engage (i.e., use EoF 4 and EoF 5), the result was many more civilian casualties than when actions stopped at EoF 2 or EoF 3. In total, this behaviour, which likely still persists, resulted in 204 civilian casualties. It also did not stop any of the seven recorded suicide attacks. Note that while we do not know the number of suicide attacks that were deterred or aborted and can therefore not make any definite
statements on the effectiveness of the military presence at the operational level, the results do indicate that the compliance-based strategy is ineffective at defending against a suicide attack that is already in progress.

How can this situation be improved? Attempts to do so first pointed out that the NATO EoF chain has originally been developed for peacekeeping operations in situations where for example a group of individuals has started throwing rocks or rioting and needs to be calmed down or dispersed. The EoF chain was never meant to provide support for assessing threats but rather aims at control and deterrence after a hostile act or clear hostile intent has already been witnessed. Bagwell (2008) has argued that the EoF chain should be replaced by the threat assessment process (TAP) shown in Fig. 3.

The TAP does give more explicit guidance than the EoF chain of Fig. 1. It decomposes the goal of threat assessment into sub-goals and provides a list of attributes and actions that can be used for making the final assessment. It does not, however, specify how attributes should be integrated, actions selected, or how other key judgments influencing the final assessment. It does not, however, specify how attributes that reduce civilian casualties in a systematic way. Thus, it is not clear how exactly to apply this process quickly in the field to make such adherence was not actually the case. The TAP may thus end up being used simply as a way of justifying decisions that were taken solely based on compliance, and avoiding being blamed for a potentially negative outcome. Can anything else be done?

6. A psychological heuristic for reducing civilian casualties in force protection problems

Franco and Montibeller (2010) distinguished between the expert and facilitated modes in OR interventions. Force protection problems are intriguing because they partly fit both modes. The problems are amenable to the expert mode because they are “real entities” and “their analysis should be objective” (Franco & Montibeller, 2010, p. 491); they are framed by NATO armed forces by using the observable quantitative metric of civilian casualties. Additionally, the problems are amenable to the facilitated mode because “clients want satisfying solutions” and “[client] participation increases commitment for implementation” (Franco & Montibeller, 2010, p. 491). We explain below why, given the data we have, a satisfying solution seems to be more feasible than a standard model. The solution itself, the process by which it was derived and the interaction with military personnel, are described afterwards; note that the descriptions are restricted due to confidentiality.

6.1. Can standard models be applied to our dataset?

Soldiers staffing checkpoints or conducting patrols in stability operations have to face what Rumsfeld (2002) poetically called “known unknowns”. The soldiers know that oncoming traffic could be hostile but they typically know very little or nothing about such hostile incidents. This is so because hostile incidents occur very rarely. This rarity also means that the probability of such incidents cannot be estimated reliably if probability is interpreted as a relative frequency. Furthermore, information recorded on hostile incidents has a very limited utility for predicting such incidents in the future. The present dataset illustrates these issues: In 1060 incident reports of situations involving motor vehicles approaching a NATO military presence between January 2004 and December 2009 in Afghanistan, there were only seven suicide attacks recorded. These seven reports discuss only the battle damage and provide little information which could be used for predicting hostile incidents in the future such as the characteristics of the hostile vehicle.

Standard frequentist decision models cannot handle this kind of data. For example, non-extreme decision trees (Breiman et al., 1984; Quinlan, 1990) cannot be induced from a dataset that has zero hostile incidents. The only frequentist model that can be induced from the dataset of 1053 nonhostile incidents is one that predicts that an approaching vehicle is never hostile and this will obviously not do. The seven incidents in which suicide attackers were missed may be added to the dataset but the problem then is that only the attribute of the number of occupants in the vehicle is available in these reports; then, the induced frequentist model would suggest that a vehicle is always hostile when there is one occupant, which would result in many false alarms and likely civilian casualties. Furthermore, it is in principle possible to construct a Bayesian model if experts were able to provide a prior probability of hostiles on a particular military presence on a particular day, but much care should be taken to ensure that the elicitation of such probabilities is reliable (Edwards & Fasolo, 2001; Katsikopoulos & Fasolo, 2006). Our search of the academic and practitioner literature and consultation with experts did not point to the existence of such subjective probabilities.

In sum, we concluded that it was not clear how to apply a standard model to our dataset. Instead, we constructed a psychological heuristic. Psychological heuristics often use non-probabilistic information, which may be easier to possess. For instance, a soldier on a
checkpoint may be able to match the name or face of a suicide attacker with intelligence information. Or an experienced soldier may have a causal theory about the characteristics and approaching behaviour of hostile vehicles. Such intuitive theories will of course not always be right but they still may be very effective, especially when they are combined with each other, as we shall see just below.

6.2. A fast and frugal tree for reducing civilian casualties in force protection problems

Fast and frugal trees have been applied to military operations already as to the problem of detecting unexploded ordnance (i.e., munitions used in war or military practice) (Fernandez, Katsikopoulos, & Shubitizde, 2010). In a dataset of 216 objects from Camp Sibert, Alabama, a fast and frugal tree outperformed Breiman et al. (1984) classification and regression trees and matched the performance of support vector machines (Vapnik, 1999), achieving 100 percent hit rate and 3 percent false alarm rate.

In Fig. 4, we present a fast and frugal tree that soldiers can use to classify oncoming traffic as hostile or not. Note that we constructed this tree during the goal-directed task-analysis before we obtained the Afghanistan reports.

During the goal-directed task analysis, we scanned the practitioner literature for mentions of attributes that could be used for predicting hostile or nonhostile and also elicited attributes via semi-structured interviews with German armed forces instructors and combat-experienced personnel. An additional round of interviews was also performed where we asked the experts their opinion about the attributes in the literature.

For example, all experts we interviewed agreed with Bagwell’s (2008) suggestion of the multiple occupants attribute as a good indicator of civilian status. This makes sense, as from the point of view of the insurgents, suicide attackers are high-value assets and thus it would be a waste of resources to place multiple attackers in a single car. The fact that all seven incident reports on successful suicide attacks on NATO forces mentioned only a single hostile killed was only discovered later.

Second, we decided on the order of the attributes and the structure of the exits at each tree level. We decided to place occupants attribute first and use it for white listing, that is, to exclude civilians rather than identity suicide attackers. White listing has been theoretically argued to increase the effectiveness of profiling against strategic attackers (Cavusoglu, Kwark, Mai, & Raghunathan, 2013). All experts agreed that this was by far the strongest attribute available and also that using a single-occupant attribute as indicating hostile status would not be very effective because it would result in many false alarms. Implementing the attribute in this way was thus the only logical choice. Furthermore, it greatly enhanced the speed of using the tree, that is, the mean number of pieces of information that need to be identified and processed across decision situations before a classification can be made. In Afghanistan, with a much lower vehicle density than the North America or Europe, most approaching vehicles will contain more than one occupant. This, in turn, means that most approaching vehicles can be classified immediately after ascertaining the value on the first attribute. This aspect of the decision aid was deemed to be a very effective stress reducer.

Next, we decided to place the compliance attribute second due to its psychological and operational importance. Note that while noncompliance in combination with single occupancy is used as threat identifier, the burden of classifying a vehicle as hostile does no longer rest wholly on the compliance attribute; an approaching vehicle, even if compliant, may now be classified as hostile if both of the other attributes are present (e.g., single occupancy plus the vehicle description matching previous intelligence reports of a suicide vehicle). This guards against cases in which suicide attackers show compliance in order to sneak up on the military presence and then detonate.

Furthermore, the focus of the tree on classifying and taking action is attuned to the job of the individual soldier caught in potential escalation-of-force situations. Unlike a senior commander in charge of multiple soldiers and checkpoints, an individual soldier is less interested in probabilities of hostile incidents and more interested in resolving a single case of an approaching vehicle as either hostile or nonhostile.

Because our tree was not fitted to the Afghanistan reports, we can apply it to this dataset and get a fair estimate of what would have happened if it had been used. First, the tree would have greatly reduced the number of civilian casualties: In the 1053 incidents, the civilian casualties would have been reduced from 204 to 78, a reduction of more than 60 percent. This result can be analysed in greater detail, as we do in Table 4. The table presents, for all 1053 nonhostile vehicles, the incorrect classifications (i.e., hostile) of the fast and frugal tree, for each type of vehicle. A vehicle's type is the combination of its attributes; for example, a vehicle can have a single occupant, be compliant and have no further threat attributes. Note that, from the perspective of the fast and frugal tree, a vehicle type is sometimes completely determined by just one or two attributes; for example, all vehicles with multiple occupants are classified as nonhostile, irrespective of the other attribute values. The table also presents the civilian casualties caused by the soldiers, also for each type of vehicle.

Table 4 shows that the tree would have greatly reduced civilian casualties for vehicles with multiple occupants, by 183 civilians. On the other hand, the tree would have substantially increased casualties for single-occupant and noncompliant vehicles, by 64 civilians. The differences would have been much smaller for single-occupant and compliant vehicles.

Because, except for the occupants attribute, the Afghanistan reports pertaining to suicide attacks did not include information on any of the other attributes used by the fast and frugal tree, its effectiveness at classifying these attacks cannot be quantified. However, it is likely that the tree would have enhanced soldier safety. As all successful attacks involved only one occupant, all suicide attackers that either behaved in a noncompliant manner or had any other threat attributes would have been correctly identified as a threat. And the tree is fast; across all 1053 incidents, only 1.2 attributes were used on average per classification and 84 percent of the time the tree could already classify a vehicle after looking up the first attribute only.

Crucially, the task of force protection is not only a task of classification (Keller, Czienskowski, & Feufel, 2014). We must also think...
about how the threat assessments of the fast and frugal tree can be used to guide action, that is, to choose which EoF levels to use. This can be done with the data in Fig. 5. For the 1053 nonhostile incidents, two measures of performance are provided for each level from EoF 2 to EoF 5: the effectiveness of a level equals the proportion of times it led to compliance and the lethality of a level is the proportion of times it led to casualties (of hostiles or nonhostiles).

As can be seen in Fig. 4, moving from EoF 3 (warning shot) to EoF 4 (disabling shot) or EoF 5 (lethal shot) increases lethality sharply with little or no increase in effectiveness. Thus if the fast and frugal tree classifies a vehicle as a nonhostile, it makes little sense to escalate higher than EoF 3. On the other hand, if a vehicle is classified as hostile, it might make sense for a soldier to start the EoF chain at EoF 3 or even EoF 4, saving valuable reaction time. A simulation-based tile, it might make sense for a soldier to start the EoF chain at EoF 3. On the other hand, if a vehicle is classified as hostile, it makes little sense to escalate higher than EoF 3.

Finally, after this research was done, we spoke about it to a number of high-level officials such as top troop psychologists and colonels. The feedback was very positive and there was an invitation to write up a best-practice report for the German Federal Armed Forces, which we did.

6.3. Conclusions and directions for future work

Even though the psychological heuristic developed here was constructed for individual soldiers facing a specific problem, it also invites an organizational and strategic perspective. One issue that immediately comes to mind is that of strategic interaction (McLay, Jacobson, & Kobza, 2005). Given an adapting and highly motivated opponent, what if the heuristic presented here is gamed, for example, by placing two suicide attackers in the car? In this case, the fast and frugal tree of Fig. 4 can be adapted to deal with the new data.

Questions of implementation are a further crucial aspect for organizational effectiveness. The adherence of the decision tree to psychological principles such as sequential information search and deterministic classification as opposed to providing probabilities as output is likely to improve adherence. The need for technological support for the application of the decision tree (e.g., telescopic sights or infrared equipment in order to better be able to ascertain the number of occupants) and its impact on effectiveness can be quantified. End-user constraints on, for example, the number of attributes that can realistically be processed or the time needed to switch between different pieces of equipment (e.g., between the primary weapon and the signal flare/dazzling laser) can be heeded in the construction phase. In sum, we believe that the use of simple and fast decision trees should be given serious consideration as a tool for assessing and responding to threat in stability operations, especially for the purpose of reducing civilian casualties.

7. What is the role of psychological heuristics in operations research?

In one of his instructive stories, Russell Ackoff (1979, p. 97) complained about pronouncing optimization models as optimal without checking if their assumptions hold: A very large intrasystem distribution problem was modelled as a linear programming problem, and its optimal solution was derived; the argument offered for implementing this solution was that its performance was superior, according to the linear programming model, to that of another solution!

Awareness of the limits of optimization is a primary argument for soft OR (Ackoff, 1979; Mingers, 2011). The competitive performance of psychological heuristics (for an example see Fig. 1) compared to standard hard OR models in problems of multi-attribute choice, classification and forecasting which is also explained by theory (summarized in Table 2), can serve as an argument for a more pluralistic approach in developing hard OR methods as well. Of course, more work needs to be done. For example, most psychological heuristics research has ignored the case of more than two alternatives or categories (for exceptions, see Hogarth & Karelaia, 2005 and Katsikopoulos, 2013), which may be more representative of real problems. But in any case, the study of psychological heuristics can serve as a conceptual bridge between soft and hard OR. This point was made in Section 2 (summarized in Table 1).

But can psychological heuristics scale up to more complex problems, as for example strategic problems with unclear objectives and multiple disagreeing stakeholders? French, Maule, and Papamichail (2009) seem to believe no when they say that psychological heuristics can be applied to “…simple decision tasks with known correct solutions” (p. 169) and to “…some tactical and operational decisions” (p. 419).

We basically agree with French et al. (2009) that it is not yet obvious how to scale up the formal models of psychological heuristics so far developed. But there are two caveats: First, psychological heuristics require only that there exists a correct solution, not that it is given to them. In fact, it has been shown that psychological heuristics perform especially well when the correct solution will be

---

Table 4
For each type of the 1053 nonhostile vehicles in the Afghanistan incidents (Wikileaks, 2010), the incorrect classifications (i.e., hostile) of the fast and frugal tree of Fig. 4, and the civilian casualties caused by the soldiers.

<table>
<thead>
<tr>
<th>Type of vehicle</th>
<th>Number of nonhostile vehicles</th>
<th>Number of civilians classified as hostile by fast and frugal tree</th>
<th>Number of civilian casualties caused by soldiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple occupants</td>
<td>890</td>
<td>0</td>
<td>183</td>
</tr>
<tr>
<td>Single occupant; noncompliant</td>
<td>74</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Single occupant; compliant</td>
<td>85</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Single occupant; compliant; further attributes</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>1053</td>
<td>78</td>
<td>204</td>
</tr>
</tbody>
</table>

---

Fig. 5. Effectiveness and lethality measures for each level from EoF 2 to EoF 5 in the NATO EoF chain, as calculated from the reports of 1053 nonhostile Afghanistan incidents (Wikileaks, 2010).
available in the future. This is a point where psychological heuristics exhibit the kind of robust power of human intuition and expertise (Klein, 1999) that is often lost in hard OR and soft OR tries to capture. The second caveat is that a psychological heuristics approach has in fact been applied to problems of understanding information about health conditions and making informed decisions about treatments. These are problems where patients, doctors, pharmaceutical companies, health administrators and policy makers often have unclear or conflicting objectives (Gigerenzer & Muir Gray, 2011). These heuristics are based on knowledge from the psychology of thinking, perception and emotion, and also from social psychology. Integrating this approach with the one presented in this chapter, and with soft and hard OR, is a key task for the future.

Acknowledgements

We wish to thank the guest editors of this special issue, two anonymous reviewers, Aris Syntetos and Özgür Şimşek for their very helpful comments on earlier versions of this manuscript.

References


Vicente, K. J. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. CRC Press.

Please cite this article as: N. Keller, K.V. Katsikopoulos, On the role of psychological heuristics in operational research; and a demonstration in military stability operations, *European Journal of Operational Research* (2015), [http://dx.doi.org/10.1016/j.ejor.2015.07.023](http://dx.doi.org/10.1016/j.ejor.2015.07.023).