Skilloscopy: Bayesian modeling of decision makers' skill

Article

Accepted Version

Skilloscopy: Bayesian Modelling of Decision Makers' Skill

Giuseppe Di Fatta, *Member, IEEE*, and Guy M^{c}C. Haworth

***Abstract*—This paper proposes and demonstrates an approach, Skilloscopy, to the assessment of decision makers. In an increasingly sophisticated, connected and information-rich world, decision making is becoming both more important and more difficult. At the same time, modelling decision-making on computers is becoming more feasible and of interest, partly because the information-input to those decisions is increasingly on record. The aims of Skilloscopy are to rate and rank decision makers in a domain relative to each other: the aims do not include an analysis of why a decision is wrong or suboptimal, nor the modelling of the underlying cognitive process of making the decisions. In the proposed method a decision-maker is characterised by a probability distribution of their competence in choosing among quantifiable alternatives. This probability distribution is derived by classic Bayesian inference from a combination of prior belief and the evidence of the decisions. Thus, decision-makers' skills may be better compared, rated and ranked.**

The proposed method is applied and evaluated in the game-domain of Chess. A large set of games by players across a broad range of the World Chess Federation (FIDE) Elo ratings has been used to infer the distribution of players' rating directly from the moves they play rather than from game outcomes.

Demonstration applications address questions frequently asked by the Chess community regarding the stability of the Elo rating scale, the comparison of players of different eras and/or leagues, and controversial incidents possibly involving fraud.

The method of Skilloscopy may be applied in any decision domain where the value of the decision-options can be quantified.

***Index Terms*—Decision making, skill evaluation, Bayesian inference.**

I. Introduction

A pilot attempts to land in marginal conditions. Multiple agencies work furiously on a major emergency. A student progresses his learning with less than total awareness, motivation or organisation. A golfer performs a 3-D rotation and translation in a powerful yet precise driver swing.

In problems of complex decision making the combined pressure of events, real-time, partial information, problem complexity and limitations on human (and computer) resources may cause the human component to take sub-optimal decisions, short of the utopian agent in the 'How To' manual. To model such systems effectively, it is necessary to model and to measure decision makers' skill.

The importance of skill evaluation and modelling of decision makers in motor and cognitive activities has been recognised in a broad range of application domains: mineral processing plant [1], surgery [2], [3], Air Traffic Controllers [4], robotic arms [5] senior police officers [6], Intelligent Tutoring System and Adaptive Learning Platforms [7], sports [8] and games [9].

Bayesian inference has been adopted successfully in many applications for model selection and decision making. Typically these methods have been applied to determine models which explain the empirical evidence of data representing best solutions to problems or data generated by complex systems.

This work proposes the application of the classic Bayes' rule to determine parametric models of decision makers' skill. Given a history of decisions in a given problem domain, can we generate a model of decision makers' skill? How close is the decision making to optimal? And more specifically, how can their competence level be rated and ranked? Finally, how can we assess, monitor and compare decision makers' skills?

Decision makers exhibit different levels of expertise and competence, which can be represented by a skill level. Probability density functions can encode uncertainties due to the lack of empirical evidence and the intrinsic variability of human factors. In this paper, we combine prior beliefs and empirical evidence of the quality of the decisions made by decision makers using a Bayesian inference process to generate a simple probabilistic model of their skill.

The only assumption underlying *Skilloscopy*, our proposed approach, is the availability of a utility function, either exact or heuristic, that provides a numerical estimation of the utility of the alternative choices. The utility function is not required to be available at the time of the decision, nor available to the decision maker. It serves as benchmark system of the decisions in their context to support the assessment process. For example, this may be the case in training and monitoring human operators, in games and sports, or in cases where the utility function is computationally prohibitive for a real time scenario. If a model and a software simulator is available, a brute force approach can provide a utility function by exploring all alternatives and evaluating their effects on the overall goal. If no explicit model of the system is available, the utility function could just be the offline assessment by an expert.

The proposed method is demonstrated and evaluated in the game-domain of Chess. Chess has always been a favourite demonstration domain in the fields of cognitive psychology and artificial intelligence as it is a well-documented, familiar, large, complex model-domain, many of whose aspects are subject to quantification. An important contribution to the study of human skills from the Chess domain is the Elo

Authors are with the School of Systems Engineering, The University of Reading, Whiteknights, Reading, Berkshire, RG6 6AX, UK, e-mail: G.DiFatta, G.Haworth@reading.ac.uk.

Manuscript received September 15, 2011. Revision received September 19, 2012.

rating system [10]. It was originally devised for Chess and adopted by the United States Chess Federation in 1960 and by the World Chess Federation (FIDE) in 1970. The Elo rating system determines the relative strength (rating) of players by an iterative inference process based on the outcome of the games. While, in general, it may be difficult to assess the accuracy of models of decision makers' in many domains, models of Chess players' skill can be compared with the Elo ratings.

This work proposes a data mining approach to model decision makers' skill and, more specifically, its contributions are summarised as follows:

- it defines a generic, domain-independent model of stochastic decision making agents;
- it applies a Bayesian inference methodology by means of an efficient adaptive algorithm to generate probabilistic models of decision makers' skill;
- it demonstrates the use of the method in the game-domain of Chess to determine players' skills from the quality of decisions (moves) rather than from game outcomes; and
- it demonstrates applications of the method to address a number of questions asked in the Chess domain.

Although the proposed method is experimentally tested on the game domain of Chess, it is not based on any specific model of the decision making process of Chess players. How a player chooses a particular move among alternative choices (variants) is not considered or modelled. The method can be easily adopted in rating skilled behaviour and general types of expertise in other domains. The method infers skills directly from the innate quality of the decisions and independently of the competitive nature of the activity. For example, this method could be effectively adopted to monitor and evaluate training and education activities.

The rest of this paper is organized as follows. Section II reviews the general approaches in Decision Theory. Section III introduces the general problem of modelling decision makers' skill and the proposed Bayesian inference method. Section IV discusses related work in skill rating, in general, and in Chess, in particular. Section V presents the application of the proposed method to generate Chess players' ratings. An experimental analysis of the method and innovative applications in the domain of Chess are presented, respectively, in Section VI and VII. Section VIII reviews related work. Finally, Section IX summarises the paper and indicates some future research directions.

II. Modelling Decision Making

Decision making under uncertainty and, more generally, Decision Theory, has been studied for a long time [11] with contributions from several academic disciplines including statistics, economics, psychology, philosophy, political and social science. Decision Theory is concerned with identifying utility values and the uncertainty of decisions. It is typically distinguished into Descriptive, Prescriptive and Normative. Normative Decision Theory is concerned with identifying optimal decisions as taken by an ideal decision maker, who is rational, fully informed and able to compute with perfect accuracy. Prescriptive Decision Theory is concerned with what people should and can do. Decision analysis [12] (for a more recent survey see [13]) is the practical application of Prescriptive Decision Theory and is aimed at finding tools and methodologies which can support people to make better decisions.

In contrast, Descriptive Decision Theory [14] is concerned with describing what people actually do. Descriptive models have been developed with the aim of capturing the underlying processes that guide human choice behaviour under uncertainty. However, the effort in understanding and describing the actual process of and reasons for decision making has not been supported by empirical studies; the development of a Descriptive Decision Theory itself has been questioned as unachievable [14].

In a long tradition, cognitive psychology and artificial intelligence have also tried to define and explain skilled behaviour such as human expertise in problem solving and decision making. The game of Chess has always been a favourite demonstration domain. Chess players' thinking [15], [16] has been studied for a long time and two main models have been provided. One mechanism is based on pattern recognition to access a knowledge database. The second approach is a search strategy through the problem space. The relative importance given to knowledge and quantitative search varies in the proposed theories of skilled behaviour in Chess [17], [18].

This work takes a descriptive data mining approach to the problem. Rather than building an explicit model of the decision making process, we propose to build a probabilistic model of the decision maker's skill, without any attempt to understand or to model the underlying decision process.

III. A Model of Decision Makers' Skill

In problems of a strategic character, human beings exhibit different levels of expertise and competence in making decisions which have a direct or indirect effect on the achievement of an overall goal. In general, we expect that the competence of decision makers is the cumulative result of abilities, training and experience, typically referred to as 'skill'.

In general, a model of decision makers' skill may be defined by a set of parameters, i.e. multidimensional skills. Without loss of generality in this work we limit our analysis to the simple model of a single parameter $c \in \mathbb{R}$. We assume that decision makers can be modelled by a numeric skill level c, which indicates their competence in solving a particular class of problems. The advantage of defining models with a single parameter is that the skill level can be directly used for ranking and rating. However, a multidimensional skill model may be more suitable to capture specific aspects in more complex domains.

Even though the average skill level of a given decision maker is expected to vary smoothly over time, decision making activities can be appropriately described as stochastic processes where decisions are generated by an agent with an apparent competence c. We introduce the concept of a stochastic reference agent $R(c)$. The agent $R(c)$, when presented with a problem with a set of alternatives, chooses by means of a stochastic process biased by its competence level.

Skilloscopy imagines that the decisions have been taken by one of a set of these stochastic agents of defined reference behaviour and skill. Skilloscopy asks the question:

> "On the assumption that the decisions observed have been made by one of the reference agents available, what is the probability associated with the hypothesis that a particular agent made the decision?"

This is a classic inverse problem which, given a set of prior beliefs, can be solved precisely by the process of Bayesian inference.

In this section a general model of decision makers' skill is introduced and a Bayesian inference method presented. The Bayesian inference of the model of decision makers' skill is defined by five components:

- a class of decision making problems,
- a utility function for the problem class,
- a reference agent with a decision likelihood function,
- a prior probability of the competence, and
- empirical/historical data, i.e. a set of problems and associated decisions of a decision maker.

Given a problem class Q, a decision making problem $q \in Q$ is associated to a set of alternatives $A_q = \{a\}$. A utility function $u(q, a)$ assigns numerical values to each feasible alternative a of the problem q.

The model is defined by the likelihood that the alternative a is chosen by a stochastic reference agent with given competence c. The likelihood depends on the particular problem q and, more specifically, on the relative utility of the alternatives A_q.

The Bayesian method is adopted to determine the model that best explains the evidence of a given decision maker's past decisions. The Bayesian inference method requires an initial prior probability, which is used to incorporate any knowledge about the probability of the model before any evidence is considered. Where no prior knowledge is available, a uniform distribution may be used.

Table I summarises the general notation used in this section.

TABLE I
GENERAL NOTATION

Notation	Description
Q	problem class
q	problem instance
A_q	set of alternatives for problem q
a	alternative
c	competence, skill level
$u()$	utility function
$L()$	likelihood function
$p()$	probability function
P_0	prior probability constant
$R(c)$	stochastic reference agent with competence c
S	set of events $\langle q, a \rangle$

A. Utility Function

The utility function provides a measure of the quality of the decisions, i.e. a benchmark system of the available alternatives.

The skill of a decision maker measures the degree of agreement of their decisions with the benchmark system.

Given a decision problem q with a finite set of alternatives A_q, the utility function $u(q, a)$ returns their utility values in an arbitrary range and some units. Where context allows, we will consider the problem q and the set of alternatives A_q implicit to simplify the notation. The utility function expresses preferences in outcomes: $u(a_1) > u(a_2)$ iff a_1 leads or is expected to lead to a better outcome than a_2.

The utility function is domain specific and its availability is the only fundamental assumption in the proposed approach. The utility function is necessary to carry out the assessment of decision makers, not to perform decision making activities. The utility function may be exact or heuristic, relative or absolute, based on an human expert, a system model or a brute-force search within a system simulator. In case the utility function is heuristic, approximations obviously introduce errors and uncertainty. The influence and sensitivity of the method to the approximation of the utility function need to be properly addressed. However this is beyond the scope of this presentation and the subject of future work.

B. Reference Agent and Decision Likelihood

The stochastic reference agent is a generic synthetic decision-maker. Given a set of alternatives and their utility values, the stochastic behaviour of the reference agent $R(c)$ is defined by the likelihood that an alternative is chosen, given its competence c.

Agents are not meant to model human decision makers. They are used to define a parametric skill model in adhering to a benchmark system.

The likelihood function $L(\cdot)$ is defined in terms of the utility function $u(\cdot)$. The likelihood function $L(a|c)$ provides the likelihood of an alternative a being chosen by the stochastic reference agent $R(c)$.

Stochastic agents $R(c)$ should cover the entire skill-range of decision makers ($c \in [0, \infty]$). The agent with zero apparent competence ($c = 0$) corresponds to making random-choices. Greater c values correspond to better competence. The ideal decision maker $R(\infty)$ always makes an optimal decision as defined by $u(\cdot)$.

The likelihood of an agent choosing an alternative a is always greater than zero, regardless of its utility value. The likelihood is a monotonically non-decreasing function of the utility ($dL/du \geq 0$) and is convex ($d^2L/du^2 \geq 0$).

The requirements for the function L are summarised as follows:

- $L(a|c)$ is finite and positive for all alternatives $a \in A_q$ and competence parameter c;
- for $c = 0$, $L(\cdot)$ is independent of $u(\cdot)$, i.e. all alternatives are equally likely to be chosen ('zero competence');
- as the competence-parameter c increases and given $u(a_i) > u(a_j)$, $L(a_i|c)/L(a_j|c)$ increases; and
- as $c \to \infty$, $L(a_i|c)/L(a_j|c) \to \infty$, i.e. less attractive options can be made arbitrarily unlikely.

Given the above requirements, we choose to define the likelihood of an alternative a being chosen by a stochastic reference agent with given competence c as

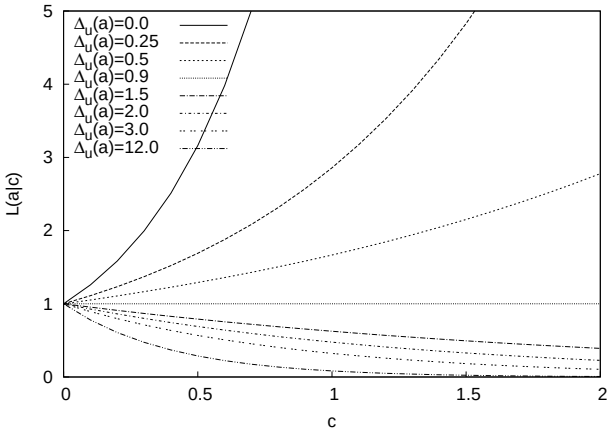


Fig. 1. The likelihood function for various Δ_u values ($k = 0.1$).

$$p(a|c) \propto L(a|c) = (u(a^*) - u(a) + k)^{-c}, \tag{1}$$

where $a^* = argmax_{a_j \in A_q}(u(a_j))$ and k is an arbitrary small constant ($k \in]0, 1]$), which ensures that $L(\cdot)$ is finite. The constant k should be small w.r.t. the typical utility values in the specific domain.

The conditional probability $p(a|c)$ of the agent $R(c)$ selecting the alternative a given the competence parameter c, is simply given by normalizing the likelihood in (1), viz.:

$$p(a|c) = \frac{L(a|c)}{\sum_{a_j \in A_q} L(a_j|c)}. \tag{2}$$

The likelihood in formula (1) is a function of the competence c and the difference ($\Delta_u(a) = u(a^*) - u(a)$) between the utility value of the alternative a and the value of the optimal alternative according to the benchmark system. $\Delta_u(a)$ is non negative by definition. Figure 1 shows an example of the likelihood as function of the competence c for a set of 8 alternatives, whose corresponding $\Delta_u(a)$ values are $\{0.0, 0.25, 0.5, 0.9, 1.5, 2.0, 3.0, 12.0\}$. Given a value $\Delta_u(a)$, agents with greater competence are more likely to identify better decisions. Given a competence value $c > 0$, worse decisions are less likely to be chosen than better ones.

The formula (1) is not the only function which can comply with the requirements discussed above. It has been chosen for its simplicity and generality, i.e. it does not depend on a specific application domain. Further investigations may be devoted to the experimental analysis of the effect of different likelihood functions in specific application domains.

In the next section we describe the Bayesian inference method to determine a probability distribution of the parameter c for decision makers from the evidence of their past decisions.

C. Inference of the parametric model

Bayesian inference [19], [20], [21] is an iterative process in which evidence modifies an initial probability distribution of belief. In each iteration, the initial distribution is the prior probability whereas the modified belief is the posterior probability.

Let us consider the event $\langle q, a \rangle$, where a is an alternative chosen for the problem q ($a \in A_q$). In the following the problem q is considered implicit to simplify the notation. The posterior probability of the parameter c, given the evidence of the choice a, depends on the *a priori* probability $p(c)$ and the conditional probability of the evidence a given the competence parameter c, as stated by the Bayes' theorem:

$$p(c|a) = \frac{p(c) \cdot p(a|c)}{\sum_c p(c) \cdot p(a|c)}. \tag{3}$$

Let us consider a set of N events $S = \{\langle q_i, a_i \rangle\}$, where $i \in [1, N]$. We iteratively apply the Bayesian rule in (3) to the set of events in S, where the *a priori* probability at step i ($i > 1$) is the posterior probability at step $i - 1$.

$$p(c|a_i) = \frac{p(c|a_{i-1}) \cdot p(a_i|c)}{\sum_c p(c|a_{i-1}) \cdot p(a_i|c)} \tag{4}$$

when considering the sequence of events $\langle q_{i-1}, a_{i-1} \rangle$ and $\langle q_i, a_i \rangle$.

In the experimental analysis we have set the initial *a priori* probability $p(c)$ to be a 'know nothing' uniform probability P_0, if not otherwise stated.

For a given set of events associated to a decision maker, the inference process produces an *a posteriori* probability distribution of the model parameter c. The expected value $\bar{c}$ is, by definition, the average apparent competence of the decision maker and can be used to generate a skill rating system. The variance of the probability distribution provides a measure of the uncertainty of the rating, similarly to [9], [22]. The uncertainty of the competence can be associated to the cumulative effect of several causes, including the uncertainty of prior belief, the intrinsic variability of human factors and the limited amount of empirical evidence. According to the central limit theorem, the variance should be inversely proportional to the square root of the sample size (empirical evidence). However, for a very large sample size we expect to find a non zero asymptotic minimum of the variance due to the intrinsic variability specific to each human decision maker.

In general, practical applications of Bayesian inference involve the subjective choice of prior probabilities and are limited by the computational complexity of numerical methods for the integration of the posterior probabilities. In the experimental analysis we investigate the effect of different prior probability distributions.

1) Adaptive algorithm: Efficient numerical methods can be used to approximate the posterior distribution arising in Bayesian inference, especially for high-dimensional functions. Methods based on Markov Chain Monte Carlo (MCMC) and Gibbs sampling have been proposed [23] and are available [24]. In our case we have adopted a simple efficient algorithm, which is briefly discussed here.

The parameter c is notionally in $[0, \infty]$ but is initialised in practical computations as in $[c_{min}, c_{max}]$, where $c_{min} \geq 0$.

We have used an adaptive detection of the range of c for a more efficient computation of the probability distribution. The parameters c_{min}, c_{max} and δ_c define a finite set of discrete values of the parameter c:

$$c_i = c_{min} + i \cdot \delta_c, \tag{5}$$

where $0 \leq i \leq (\frac{c_{max}-c_{min}}{\delta_c})$.

The three parameters are adjusted during execution to allow a better resolution of the distribution of c. The iterative process starts from a wide range $[c_{min}, c_{max}]$ with a coarse precision ($\delta_c = 0.1$). At each iteration step, the range is narrowed and the precision increased (δ_c is decreased). This results in a more efficient computation in terms of runtime and memory requirements.

IV. Skill Rating

Many rating systems in games and sports are based on the Bradley-Terry model for paired comparisons [25]. The assumptions of such rating systems are that the strength of a player can be described by a single value (rating) and that expected game results depend only on the difference between the ratings of the two players.

Ratings based on pairwise comparisons perform an indirect inference of the skill level by means of the outcomes of competitive activities involving two or more individuals. Such rating systems are intrinsically relative. Most rating systems fall into this category, including the most prevalent, the Elo system [10].

In comparison, *Skilloscopy* enables the direct measurement of skill and could be applied, more generally, in non-competitive domains of complex decision making where professional standards must be maintained despite the pressures of events, time constraints, partial information, problem complexity and ability.

A. Skill Rating in Chess

The Elo system [10], [26], perhaps the best known rating system, was originally created for Chess and later adapted to other games, video games and sports. For example, the U.S. Table Tennis Association have adopted it to rate players [26]. The National Collegiate Athletic Association (NCAA) uses several rating systems, including Jeff Sagarin's computer ratings [27]. Sagarin's ratings are used for NCAA basketball teams and in the calculation of the Bowl Championship Series (BCS) computer ratings in college football. Sagarin's overall ratings are based on two different ratings, one of which is a modification of the Chess Elo rating system.

Within a pool of players, Elo differences are meaningful, but Elos from different pools of players are not comparable as Elos have no absolute meaning. They are also affected by the Elos being imported (exported) to (from) the pool by players entering (leaving). The Elo scales for human players and for computers [28] are said to have been affected by both deflationary and inflationary forces [29]. It is possible for the same ELO figure to be attained in a different era by a player who in fact, in absolute terms, plays worse (inflation) or better (deflation).

Elo ratings are determined from the results of games and not by the innate quality of the moves played: they therefore measure competitive performance rather than underlying skill. There have been criticisms of the Elo approach [30] and improvements [29], [9], [31], [22] have been proposed. However, they are still based on the outcomes of paired comparisons and affected over time by the changing player population. As such, these approaches cannot accurately determine:

- inflationary trends in ratings, i.e. changing the quality of play at a specific Elo figure,
- the relative skill of players in a specific part of the game, e.g. the 'opening' or 'endgame',
- the relative skill of contemporary players in different leagues and/or of different eras,
- whether a match or game was won by good play or lost by bad play, or
- whether a player is playing abnormally well and perhaps cheating.

In contrast, a few systems have been proposed to assess, rank and rate absolute Chess skill. Authors in [32] and in [33] use the 'error' of move-decisions to calculate mean-error, but they do not use the full move-context of the decisions, nor do they use a Bayesian approach.

V. A Bayesian Inference of Chess Players' Rating

The application of the proposed method to the game domain of Chess and, in particular, to the problem of Chess players rating is presented. Chess players, at their turn of play, face decision making problems over the set of legal moves for the given position on the board. Their decisions may often be seen as sub-optimal, if compared to a good benchmark system, e.g. the analysis of alternatives provided by a better player.

In the remainder, we show how the general approach of Section III can be specialised for the domain of Chess. Table II summarises some additional notation used in this Section.

TABLE II
Additional notation for Chess domain

Notation	Description
CE	Chess Engine, a computer programme
W	white
B	black
d	number of plies, a Chess Engine parameter
n_{mv}	number of move variants, a Chess Engine parameter
m	Chess move
v	utility values associated to a Chess move
q	a problem, a Chess board position
M_q	set of candidate moves in a given Chess position q
V_q	set of pairs $\langle m, v \rangle$, Chess move and its utility value
N	number of board positions (problems)
$\mathbf{G}$	set of Chess games
$S_{0,i}$	set of events associated with a lost game
$S_{1,i}$	set of events associated with a won game
$S_{\frac{1}{2},i}$	set of events associated with a drawn game
ELO	FIDE Elo rating of a Chess player

A. Reference Chess Engine and Stochastic Chess Agent

During a Chess game players make decisions according to individual judgement under time pressure. Rational decision making requires a definite set of alternative actions and knowledge of the utilities of the outcomes of each possible

action. A player's skill is a measurement of their ability to make choices as close as possible to the optimal ones. In order to assess a player's skill level, we ideally need the 'best move' benchmark but this is only available via Endgame tables in the Endgame Zone. In the general case of the whole game this is clearly not feasible. However, significant advances have been made in the last decades in terms of Chess engines' playing strength [28]. Chess engines can be adopted as domain experts to provide utility values for alternative moves.

Given a Chess Engine CE, a Chess board position analysis results in a list of recommended 'best' moves and their heuristic utility values in pawn units (centipawns). The value associated with a move corresponds to the estimated advantage of the position that the move will lead to after a number of turns. Because of time constraints and the exponential nature of the computational complexity, the analysis of the Chess engine can only be performed up to a given maximum depth d (number of plies) and for a limited number n_{mv} of alternative variants.

For the purpose of this work we consider the reference Chess engine $CE(d, n_{mv})$.

In contrast to the ideal Endgame Zone scenario, three main factors introduce an approximation in the evaluation of a Chess position in terms of candidate moves and relative values. They are the limited search depth and span (parameters d and n_{mv}) and the heuristic nature of the Chess engine's analysis. This is a general problem of sensitivity to heuristic utility functions in Descriptive Decision Theory. The influence of this approximation in our analysis will be the scope of future investigations.

The analysis of a position q is a function f_{CE} which provides a list $V_q = \{(m_i, v_i)\}$ of candidate moves m_i and their estimated utility values v_i ($1 \leq i \leq n_{mv}$):

$$f_{CE} : q \rightarrow V_q.$$

Let $M_q = \{m_i\}$ be the set of candidate moves in V_q and $v_{max} = max_i\{v_i\}$.

To model human players' skill we associate a likelihood function L (Equation 1) with the reference Chess engine CE to create a stochastic Chess agent $R(c)$. The stochastic Chess agent will not always play the best move; it uses a likelihood function to select one of the alternatives provided by the reference Chess engine.

Following the approach in section III-B the probability of $R(c)$ selecting a move m_j is given by:

$$p(c|m_j) = \begin{cases} \frac{(v_{max}-v_j+k)^{-c}}{\sum_{m_i \in M_q}(v_{max}-v_i+k)^{-c}}, & \text{if } m_j \in M_q; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

In this case, we have generalised to feasible moves which may not be considered by the reference Chess engine due to the limited number n_{mv} of alternative variants.

VI. Experimental analysis

For the experimental analysis the following resources have been used:

- publicly available data in Portable Game Notation (PGN) from sources including the ChessBase database [34],
- TOGA II v1.3.1 [35], a publicly available, reputable and widely used Chess engine,
- the Universal Chess Interface (UCI) protocol [36] for Chess engine input/output, and
- Java code implementing UCI and the Bayesian inference method (III-C).

During a preprocessing phase positions were acquired from the games, ignoring the first 12 moves by each side (assumed to be 'from the book'). Configurations of the Chess board were converted to a set of events $\{e = (p, m)\}$, where each move m is made in a board position p. Positions were analysed using the Chess engine, which provided the utility values for a finite set of alternatives moves.

Each analysis was carried out to depth $d = 10$ plies. The Chess engine was configured to determine and report the top moves ($n_{mv} = 10$) it found in each position. Both values were chosen as compromises between computing speed and comprehensiveness of the data. Depth 10 is not considered sufficient to outplay the stronger players in our samples, but apparently it suffices to identify most of their inaccuracies.

Finally, the Bayesian inference process has been applied to the preprocessed data to generate the probability distribution of the parameter c.

In all our tests the arbitrary constant k of formula (1) has been set to 0.1, which corresponds to a thousandth of the value of a pawn.

A. Composite reference Elo players

This experiment shows that the proposed Bayesian approach is able to detect different skill levels among players with different Elo ratings.

The decisions of players of different skill levels have been analysed. We have used all available 3432 games in which both players had Elo ratings within 10 points of some Elo figure, e.g. games of players rated between 2390 and 2410. Games were grouped according to the Elo rating of the players. Each group contains games between players with a similar Elo rating ($ELO_{min} \leq ELO(player) \leq ELO_{max}$). The number of games and the number of positions (move-events) which have been included in the datasets of composite reference players are given in Table III.

We have applied the Bayesian inference method described in section III-C to each dataset of Table III independently. The probability distributions of the parameter c is shown in Figure 2(a). A summary of these distributions is provided in Table III in terms of the mean, standard deviation and 95% credibility region (CR) for c.

The expected value of c, $E(c) = \bar{c}$, measures the average quality of moves played in the games. We refer to $\bar{c}$ as the 'apparent skill'. The standard deviation σ_c measures the uncertainty of the apparent skill level, caused by the varying performance of the players, the finiteness of the data and the 'spread' of initial belief.

As expected, the standard deviation and the width of the credibility region depend on the amount of input data and

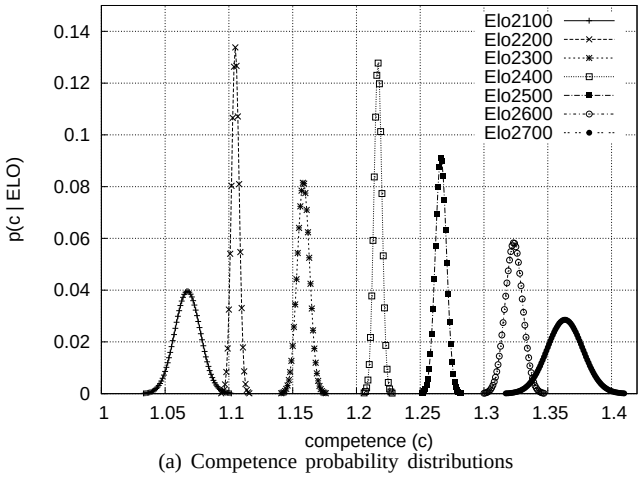
TABLE III
CHESS GAMES DATASETS

Player	ELO_{min}	ELO_{max}	Period	Games	N	CR_{min}	CR_{max}	$\bar{c}$	σ_c	$\sigma_c \cdot N^{\frac{1}{2}}$
Elo2100	2090	2110	1994-1998	213	11727	1.047	1.089	1.068	0.0101	1.091
Elo2200	2190	2210	1971-1998	2771	140390	1.098	1.112	1.105	0.0030	1.117
Elo2300	2290	2310	1971-2005	1064	57627	1.147	1.169	1.158	0.0049	1.171
Elo2400	2390	2410	1971-2006	3055	152589	1.210	1.224	1.217	0.0031	1.219
Elo2500	2490	2510	1995-2006	1646	83748	1.256	1.276	1.266	0.0044	1.266
Elo2600	2590	2610	1995-2006	746	37623	1.309	1.337	1.323	0.0068	1.326
Elo2700	2690	2710	1991-2006	121	9279	1.335	1.391	1.364	0.0140	1.345

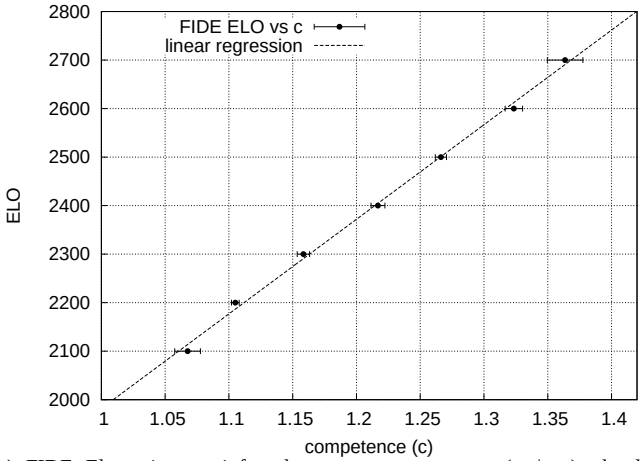
is broadly proportional to $\frac{1}{\sqrt{N}}$, where N is the number of positions. The $Elo2100$ and $Elo2700$ datasets contain less data than the others and so show slightly higher standard deviation. In the next section we analyse the effect of the amount of input data in more detail.

In Figure 2(b) a linear regression fit of the apparent skill of composite reference Elo players is shown. This shows that the proposed skill rating system correlates closely with the FIDE Elo rating system in spite of the fact that they use different information to infer the rating. The apparent skill is based on the utility of individual moves while FIDE ELO is based on the results of whole games.

The parameters of the linear regression model can be used to convert the apparent competence $\bar{c}$ into an equivalent Elo rating ($c2ELO$), i.e. $c2ELO = 1949.53 \cdot \bar{c} + 32.39$.



(a) Competence probability distributions



(b) FIDE Elo rating vs inferred apparent competence ($\bar{c} \pm \sigma$): the linear regression model is $y = 1949.53 \cdot x + 32.39$.

Fig. 2. Posterior probability distributions of the model $R(c)$ for composite reference Elo players.

B. Convergence analysis

In order to check the derivation process of the probability distribution of the apparent skill, we have taken snapshots at different iteration steps (i.e. number of positions). Figure 3 shows the analysis that has been carried out on the 2400 Elo data. The curves in Figure 3(a) show the evolution of the probability distribution during the refinement of the Bayesian inference process. The expected value $\bar{c}$ (Figure 3(b)) quickly converges and the standard deviation (Figure 3(c)) decreases as the inference process draws on more data. The asymptotic value of the standard deviation is a measure of the intrinsic variability of the skill level.

C. Skill difference in players with similar Elo ratings

In this section we present the experimental test aimed at investigating differences of apparent skill in single games between players with a similar Elo rating. Note that such a difference cannot be detected by the Elo system in principle. The Elo rating captures an average performance of a player in terms of game outcomes and not in terms of the quality of the moves.

Given a set of games $\mathbf{G}$ among players with similar Elo rating (e.g. E2400), from each game we have extracted two sets of events, one for each player, white (W) and black (B). The events are associated with the outcome of the game ($loss = 0, win = 1, draw = \frac{1}{2}$). The events are aggregated by outcome to generate three sets of events S_0, S_1 and $S_{\frac{1}{2}}$. The set S_1 contains sets of move-events which have been made by players who won the game, S_0 those made by players who lost and $S_{\frac{1}{2}}$ those made by players who drew.

$$\mathbf{G} \rightarrow \begin{cases} S_{0,W} \cup S_{0,B} \rightarrow & S_0 = \{S_{0,i}\}, i = 1..n_0 \\ S_{1,W} \cup S_{1,B} \rightarrow & S_1 = \{S_{1,i}\}, i = 1..n_1 \\ S_{\frac{1}{2},W} \cup S_{\frac{1}{2},B} \rightarrow & S_{\frac{1}{2}} = \{S_{\frac{1}{2},i}\}, i = 1..n_{\frac{1}{2}} \end{cases}$$

Each set $S_{r,i} = \{\langle q, m \rangle\}$, where $r \in \{0, 1, \frac{1}{2}\}$, contains the moves of a single player during a single game.

We have applied the Bayesian inference process to each set of events $S_{r,i}$ and compute $\bar{c}$ for each of them.

In this case, the apparent skill $\bar{c}$ measures the quality of moves played by a 'single' player during a 'single' game,

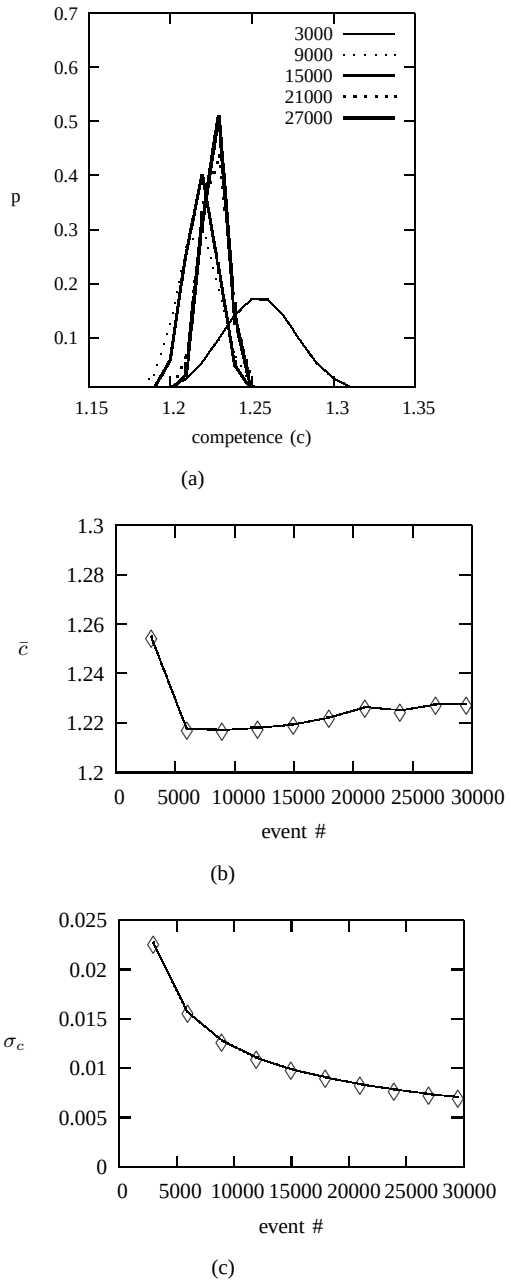


Fig. 3. Convergence evolution for the dataset 2400 Elo: a) *probability distribution at different iteration steps*, b) *expected value*, c) *standard deviation*.

with a consequent expected high uncertainty because of the limited amount of data on which the inference is carried out.

We have computed first order statistics of the apparent skill $\bar{c}$ over the three sets S_0, S_1 and $S_{\frac{1}{2}}$. In this test we have used 602 games of the dataset E2400, 313 of which were a win and 289 a draw ($n_0 = n_1 = 313$ and $n_{\frac{1}{2}} = 578$). The average apparent skill $\mu_{\bar{c}}$ over all 1204 $S_{r,i}$, regardless of the result r, is 1.2109. The results over each set S_0, S_1 and $S_{\frac{1}{2}}$ are shown in Table IV.

In spite of the small number of events in a single game the Bayesian approach is able to detect a meaningful difference between the two opponents of a game having similar Elo ratings. On average, players who have won the game have a higher apparent skill $\bar{c}$ than their opponents who have lost. Players who have drawn have even higher apparent skill. This can be explained considering that in drawn games both opponents have played well with no or irrelevant errors. In draws the intrinsic quality of the game is in general higher. When a player has reached a significant advantage during the game, they may prefer to play an easy and safe strategy. They can even afford to make small errors provided the outcome is ensured. In this case there is a lack of motivation to play high risk tactics even if optimal.

TABLE IV
ANALYSIS OF THE OPPONENT PLAYERS IN THE DATASET E2400

set	n_r	$\mu_{\bar{c}}$	$\sigma_{\bar{c}}$
S_0	313	1.1493	0.0686
S_1	313	1.2302	0.0623
$S_{\frac{1}{2}}$	578	1.2339	0.0460

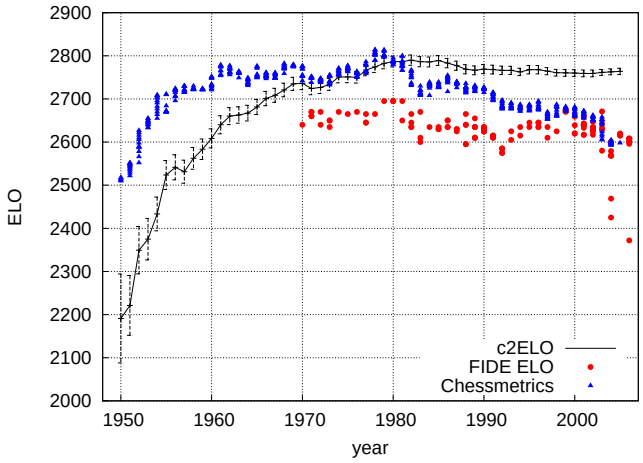


Fig. 4. Viktor Lvovich Korchnoi (games from 1950 to 2006)

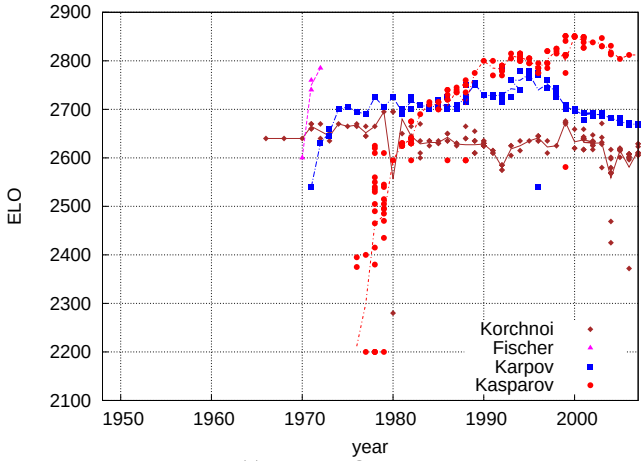
VII. APPLICATIONS

In this section we demonstrate some interesting applications of the proposed method in the domain of Chess.

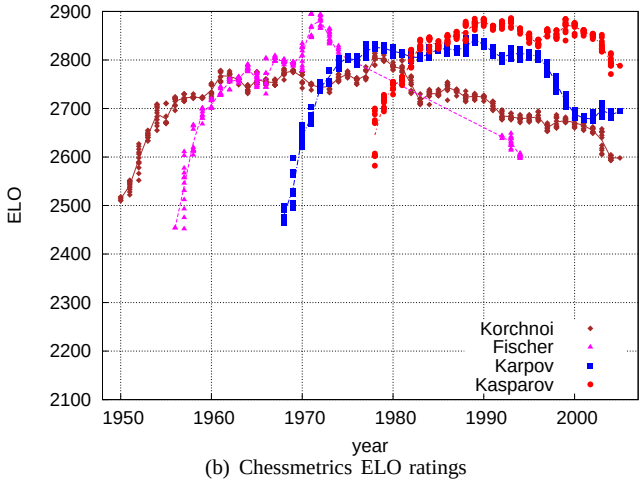
In the first example, the method is used to generate the skill profile of players over several decades, even before the official adoption of the Elo rating system. We compare the profile of a top player with the official FIDE Elo ratings and with the ratings generated by Chessmetrics [37]. FIDE have published players' Elo ratings every three months since 1970. Chessmetrics has been chosen for comparison because it is an attempt to improve the accuracy of the statistical inference method of the Elo system. It has also been applied to game data before the FIDE adoption of the Elo system. Both Elo and Chessmetrics use only the results of games (paired comparisons) to infer players' strength.

We have also generated a historical comparison of a few top players' profiles. This scenario is used to to carry out an experimental analysis of the sensitivity to the prior probability.

In the second example, we present a chart that is suitable to visualise the within-game skills of players and their opponents. In particular, we have generated this chart to analyse an aspect



(a) FIDE ELO ratings



(b) Chessmetrics ELO ratings

Fig. 5. Selected top players' ratings based on paired comparisons

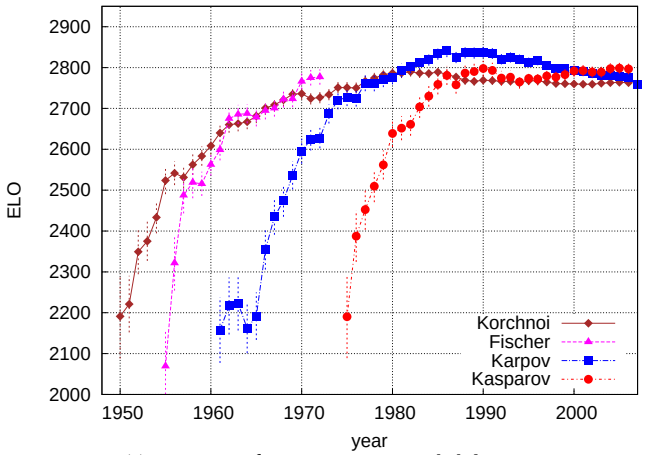
of the famous and controversial [38] final stage of the 1948 World Championship.

Finally, in the third example, we show how to use ratings and profiles generated by the proposed method to analyse the performance of players accused of cheating.

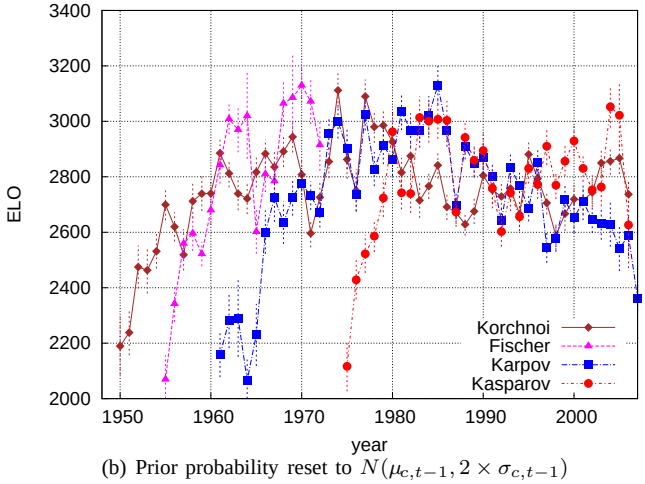
A. Profiling Human Skills over Time

We have selected all the publicly available games of a few top players to generate and compare their skill profiles over many years. The apparent competence $\bar{c}$ has been converted into an equivalent Elo rating (*c2ELO*) by using the regression model obtained in section VI-A Figure 2(b).

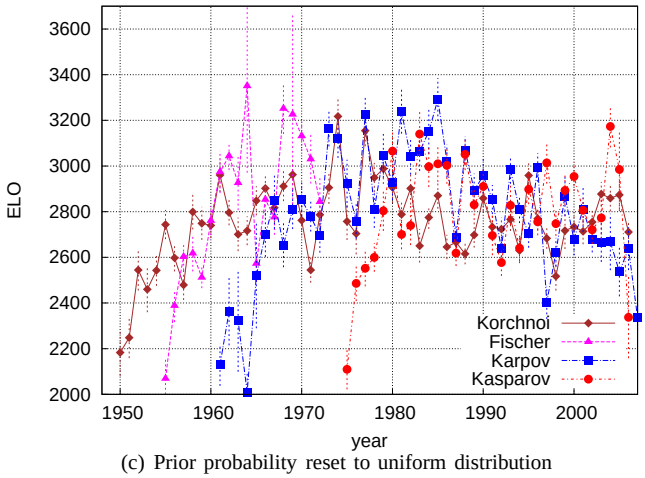
First we have generated the skill profile of Viktor Lvovich Korchnoi (1931-) to compare the proposed method to other Chess rating methods. Korchnoi is currently the oldest active grandmaster and he is considered one of the strongest players who never won the World Championship (WC). He played a candidate final (1975) and two WC finals (1978, 1981). His longevity near the top of the international Chess competitions and the turbulent political environment in which he played (he defected from the U.S.S.R in 1976) make the study of his historical ratings particularly interesting.



(a) Running inference: no prior probability reset



(b) Prior probability reset to $N(\mu_{c,t-1}, 2 \times \sigma_{c,t-1})$



(c) Prior probability reset to uniform distribution

Fig. 6. Historical comparison of top players (c2ELO ratings) over time: three different initialisations of the prior probability at the beginning of each year

The chart in Figure 4 shows the equivalent Elo rating (*c2ELO*) obtained with the proposed method, the actual Elo rating extracted from game annotations, the average Elo rating per year, the Chessmetrics rating and the average Chessmetrics rating per year. Korchnoi's *c2ELO* profile (Figure 4) shows the consistency of the proposed method with other methods based

on paired comparisons. Inference based on paired comparisons may be affected by the frequency of played games, the number and strength of similar and better opponents and other factors. The rating provided by the proposed Bayesian inference is based on the intrinsic quality of individual moves.

In Figure 4 two synchronous drops in Korchnoi's Elo and Chessmetrics ratings are evident in the periods 1971-72 and 1980-83. They indicate that Korchnoi was not performing as before in terms of game outcomes.

The rating produced by the Bayesian inference is consistent with the other two methods before 1980. In comparison, in the period from 1980 to 1986 the Bayesian method indicates a stable competence. In this period he may have lost games against better opponents, but he did not show a worse competence in terms of game quality. This may be related to the dominance of other players, e.g Karpov and Kasparov (see 6(a)).

In the previous section the excellent linear regression fit in Figure 2(b) has shown that the Bayesian inference method correlates well with the FIDE Elo rating system. Nevertheless, the evident inconsistency in Korchnoi's ratings in the period 1980-86 may suggest that the Bayesian inference method is able to detect the difference between a game lost against a better opponent and a game lost because of a poorer competence than in the past.

For a longitudinal comparison of players we have selected the players awarded the World Champion title between 1970 and 1990: Bobby Fischer (1943, 2008), Anatoly Karpov (1951-) and Garry Kasparov (1963-). We have also included Korchnoi in the comparison. In the analysis for these players we have used all available games, a total of 9404 games and 221965 positions.

Figure 5 reproduces the FIDE Elo ratings and Chessmetrics ratings [37] of the selected players.

Figure 6 show a comparison of the four players selected. In this case, we have carried out three different Bayesian inferences varying the initialisation policy of the prior probability. Since the analysis covers several decades we have decided to reset the prior probability at the beginning of each year. In the first year (1950) the uniform probability is always used.

Figure 6(a) reproduces the profiles when a continuous inference is performed: the prior probability at the beginning of the year is the posterior probability of the previous year.

Figure 6(b) reproduces the profiles when the prior probability at the beginning of the year is set to $N(\mu_c, 2 \cdot \sigma_c)$, where μ_c and σ_c are the statistics of the posterior probability of the previous year.

Figure 6(c) reproduces the profiles when the prior probability at the beginning of the year is set to a uniform probability: each year is analysed independently.

The different initialisations of the prior probability correspond to a different weighting of the knowledge w.r.t. new empirical evidence. At the beginning of each year the prior knowledge is provided by the inference over the previous year. The three initialisations methods correspond to a full propagation, partial propagation and no propagation of this knowledge.

As expected when a continuous inference is performed the curves are smooth, as the past knowledge works as low-pass filter. When prior knowledge is given less or no weight, high frequency components are not filtered and sudden variations of the skill are shown.

B. Within-game Skill Chart

Each Chess game provides two lists of moves, one for each players, and is associated with a result. Each list of moves can be used to generate apparent competences ($\bar{c}$) for the two opponents during the game. The within-game skill chart can be used to convey information about the performance of two opponents of games and their outcomes.

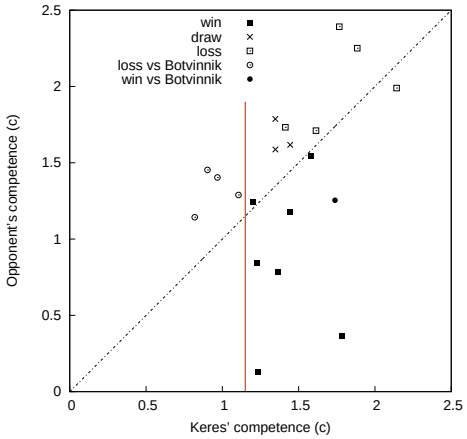


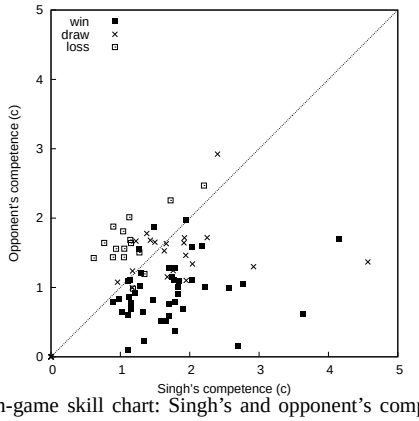
Fig. 7. Within-game skill chart: Keres' and opponent's competence ($\bar{c}$) over single games (World Championship 1948)

An example of this type of chart is shown in Figure 7, which is based on Keres' games at the 1948 World Championship. Each data point represents a game between Keres and his opponent and is shown with a symbol to indicate win, loss or draw for Keres. The horizontal axis is associated to Keres' average competence during the game; the vertical axis with his opponent's average competence.

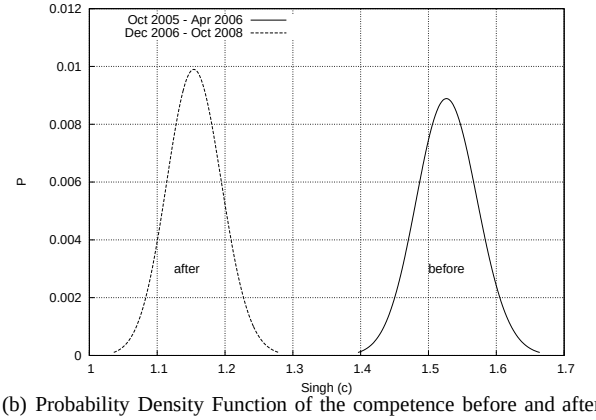
For example, the solid square at coordinates $(1.8, 0.3)$ in the chart of Figure 7 means a victory for Keres: in this game Keres made moves with an average competence $\bar{c} = 1.8$ and his opponent with $\bar{c} = 0.3$. In this game, Keres won easily against an opponent who performed poorly.

Keres' World Championship performance against Botvinnik in 1948 has long been a matter of speculation, as it is rumoured that he was under pressure not to impede the latter's progress to the title. The event was played as a quintuple round robin. Keres' and opponents' competence per game have been computed for the 20 games in which he was involved (Figure 7). The chart has symbols (cross, square and circle) to represent the outcomes of games; in this particular chart there are also two additional symbols (circle and filled circle) to identify the games Keres played against Botvinnik. Keres lost the first four games against Botvinnik and won the last only when Botvinnik had already secured the title. The vertical line at about $x = 1.1$ corresponds to Keres' competence $\bar{c} = 1.1$. The line clearly separates the first four games Keres played against Botvinnik from all his other games in the championship. In those four games Keres' competence c is below 1.1; in all other games it is above 1.1. The chart shows

that in those four games Keres clearly performed worse than in any other game of the competition. If this was intended or not is, of course, out of the scope of the analysis. Nevertheless, such a tool may be used to support or to reject hypotheses and to motivate further analysis and, eventually, investigations.



(a) Within-game skill chart: Singh's and opponent's competence ($\bar{c}$)



(b) Probability Density Function of the competence before and after allegations

Fig. 8. D. P. Singh's (games from Oct. 2005 to Oct. 2008)

C. Alleged Chess Cyborgs

There have been quite a number of players suspected of fraudulently receiving computer advice during play (e.g., [39]).

D.P. Singh's play came under suspicion in the second half of 2006 [40]. We have analysed all available games from Oct. 2005 to Oct. 2008 (Figure 8) over the entire period and over the two periods before and after the allegations.

Figure 8(a) shows the within-game chart for all games over the entire period. In some games he played at an exceptionally high skill level.

The two apparent competence profiles before and after the allegations are shown in Figure 8(b). These profiles are well separated and indicate a drop of skill level after the allegations were made.

These example suggest that the proposed method could be useful to create real-time skill monitoring applications to deter clandestine activity and to help focus the Tournament Director's finite forensic resources appropriately during play.

VIII. Related Work

This work generalises and extends in several ways the work in [41], [42], [43].

The work in [41] introduced the concept of reference agents for the Endgame Zone (EZ), defined as that part of Chess for which Endgame Tables (EGTs) have been computed. An EGT provides the value, win/draw/loss, of a position and its depth to some 'win goal' (e.g., to mate) if decisive: Chess engines play optimal moves in EZ by using EGTs.

The work in [42] introduced the concept of the reference agent based on a Chess engine. The work in [43] presented a Bayesian inference method to generate Chess players' ratings and provided the first experimental analysis. The present work extends the analysis of Chess players' skill in [43] in several ways including the following. A small-error linear regression fit shows that the method provides the same discriminative power as indirect statistical inference methods based on paired comparisons such as the Elo rating system. An experimental analysis of sensitivity to prior probabilities is carried out. A number of interesting and novel applications of the method to the game of Chess are demonstrated.

IX. Conclusions

The problem of modelling decision makers's skill has been investigated. The proposed approach, *Skilloscopy*, is based on the definition of a general stochastic model and a Bayesian inference method. It does not assume any domain-specific model of the decision making process.

The approach has been demonstrated in the game-domain of Chess. The experimental analysis has shown the viability of rating players' skill by benchmarking against Chess engines. The statistical inference is based on the quality of decisions, rather than on paired comparisons (game outcomes) as in previous approaches.

The method has been successfully applied to a large set of Chess game data and validated with the FIDE Elo ratings. The experimental analysis has provided evidence of the accuracy of the method in estimating the skill level of players regardless of the outcome of the games and of the opponent rating.

Further work will address the generalization of the method to multidimensional skills and the influence of approximations of the utility function.

In principle, the proposed method can be effectively adopted in similar domains, where an accurate method to determine utility values is available. It can be used, for example, to analyse in real-time the likely abilities of students and skilled workers in defined-process scenarios.

References

[1] O. Haavisto and A. Remes, "Data-based skill evaluation of human operators in process industry," in *Control Automation and Systems (ICCAS), 2010 International Conference on*, Oct. 2010, pp. 707 –712.

[2] M. Aizuddin, N. Oshima, R. Midorikawa, and A. Takanishi, "Development of sensor system for effective evaluation of surgical skill," in *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, Feb. 2006, pp. 678 –683.

[3] E. Lorias, M. Minor, S. Ortiz, P. Olivares, and J. Gnecchi, "Computer system for the evaluation of laparoscopic skills," in *Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, Sept. 28 - Oct. 1 2010, pp. 19 –22.

[4] K. Yacef and L. Alem, "Evaluation of learner's skills in the context of dynamic and complex systems," in *Systems, Man, and Cybernetics, 1996., IEEE International Conference on*, vol. 3, Oct. 1996, pp. 2200 –2204 vol.3.

[5] A. Koivo and D. Repperger, "Skill evaluation of human operators," in *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on*, vol. 3, Oct. 1997, pp. 2103 –2108 vol.3.

[6] R. Hartley and G. Varley, "The design and evaluation of simulations for the development of complex decision-making skills," in *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, 2001, pp. 145 –148.

[7] G. M. Goh, C. Quek, and D. Maskell, "Epilist II: Closing the loop in the development of generic cognitive skills," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 40, no. 4, pp. 676 –685, July 2010.

[8] K. Watanabe and M. Hokari, "Kinematical analysis and measurement of sports form," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 36, no. 3, pp. 549 –557, May 2006.

[9] R. Herbrich, T. Minka, and T. Graepel, "$TrueSkill^{TM}$: A Bayesian skill rating system," in *Advances in Neural Information Processing Systems (NIPS 2006)*. MIT Press, 2007, pp. 569–576.

[10] A. Elo, *The Rating of Chessplayers, Past and Present*. Arco, ISBN 0-668-04721-6, 1978.

[11] D. North, "A tutorial introduction to decision theory," *Systems Science and Cybernetics, IEEE Transactions on*, vol. 4, no. 3, pp. 200 –210, Sept. 1968.

[12] H. Raiffa, *Decision Analysis: Introductory lectures on choices under uncertainty*. Reading, MA: Addison Wesley. Republished by McGraw-Hill, 1968.

[13] Ward, Ralph and von Winterfeldt, Ed., *Advances in Decision Analysis - From Foundations to Applications*. Cambridge University Press, 2007.

[14] S. Dillon, "Descriptive decision making: comparing theory with practice," in *33rd Annual Operational Research Society of New Zealand Conference*, Aug. 1998, pp. 99–108.

[15] A. De Groot, *Het denken van den schaker*. Amsterdam, Noord Hollandsche, 1946.

[16] ——, *Thought and choice in chess (2nd ed.) (Revised translation of De Groot, 1946)*. The Hague: Mouton Publishers, 1978.

[17] F. Gobet, "Chess players' thinking revisited," *Swiss Journal of Psychology*, vol. 57, pp. 18–32, 1998.

[18] F. Gobet and N. Charness, *Expertise in chess, Chess and games. Cambridge handbook on expertise and expert performance*. Cambridge, MA: Cambridge University Press, 2006.

[19] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis (2nd ed.)*. Springer-Verlag, 1985.

[20] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, (2nd ed.)*. Chapman and Hall/CRC, London, 2003.

[21] P. M. Lee, *Bayesian Statistics: An Introduction (3rd ed.)*. Wiley, 2004.

[22] R. Coulom, "Whole-History Rating: A Bayesian rating system for players of time-varying strength," in *Proceedings of the Conference on Computers and Games*, Beijing, China, 2008.

[23] T. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[24] J. K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial Introduction with R and BUGS*. Academic Press / Elsevier, 2011.

[25] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs. I. The method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[26] M. E. Glickman, "A comprehensive guide to chess ratings," *American Chess Journal*, vol. 3, pp. 59–102, 1995.

[27] B. West, "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament," *Journal of Quantitative Analysis in Sports*, vol. 2, no. 3, Article 3, 2006.

[28] The Swedish Chess Computer Association. The SSDF rating list. [Online]. Available: http://ssdf.bosjo.net/list.htm

[29] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 48-3, pp. 377–394, 1999.

[30] J. Beasley, *The Mathematics of Games*. Dover ISBN 0-4864-4976-9, 2006.

[31] P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel, "$TrueSkill$ Through Time: Revisiting the History of Chess," in *Advances in Neural Information Processing Systems (NIPS 2007)*. MIT Press, 2008, pp. 931–938.

[32] M. Guid and I. Bratko, "Computer analysis of world chess champions," *ICGA Journal*, vol. 29-2, pp. 65–73, 2006.

[33] C. Sullivan. Comparison of great players, 2008. [Online]. Available: http://www.truechess.com/web/champs.html

[34] ChessBase GMBH, Mexikoring 35, D22297 Hamburg, Germany. Chessbase player database. [Online]. Available: http://www.chessbase.com

[35] T. Gaksch. Toga II 1.3.1 Chess Engine. [Online]. Available: http://www.superchessengine.com/togaii.htm

[36] S. Meyer-Kahlen. (2004, Apr.) Definition of the Universal Chess Interface. [Online]. Available: http://wbec-ridderkerk.nl/html/UCIProtocol.html

[37] Sonas, J. (2011, Sept.) Chessmetrics. [Online]. Available: http://www.chessmetrics.com

[38] C. C. Moul and J. V. Nye, "Did the Soviets Collude? A Statistical Analysis of Championship Chess 1940-78," *Journal of Economic Behavior and Organization*, vol. 70, no. 1-2, pp. 10–21, 2009.

[39] F. Friedel, "Cheating in chess," in *Advances in Computer Games 9*, 2001, pp. 327–346.

[40] Chessbase News. (2007, Jan.) D.P. Singh: Supreme Talent or Flawed Genius? [Online]. Available: http://www.chessbase.com/newsdetail.asp?newsid=3595

[41] G.McC. Haworth, "Reference fallible endgame play," *ICGA Journal*, vol. 26-2, pp. 81–91, 2002.

[42] ——, "Gentlemen, stop your engines!" *ICGA Journal*, vol. 30-3, pp. 150–156, 2007.

[43] G. Di Fatta, G. Haworth, and K. Regan, "Skill rating by Bayesian inference," in *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, Apr. 2009, pp. 89 –94.

Giuseppe Di Fatta has been a lecturer in Computer Science at the University of Reading (UK) since 2006. In 1999 he was a research fellow at the International Computer Science Institute (ICSI), Berkeley, California (USA). In 2000-04 he was with the High Performance Computing and Networking Institute of the National Research Council (Italy). In 2004-06 he was with the University of Konstanz (Germany). His research interests include data mining, scalable algorithms, distributed and parallel computing and multi-disciplinary applications. He has published over 60 papers in peer-reviewed conferences and journals. He has organized and chaired international workshops and conferences in Data Mining, Distributed Systems and Computer Networks. He has been a member of the IEEE since 2002.

Guy Haworth has been a lecturer at the University of Reading since 2003. He has an MA (Maths) from Oxford University, a Diploma in Computer Science from Cambridge University (1969), and researched the issues of performance, parallelism and integrity of real-time systems at Cambridge until 1972. In 30 years in industry, his roles ranged over product development, sales and marketing, customer service and consultancy, mainly for International Computers Limited (ICL). His focus was Information Management and latterly Enterprise Architecture. He has published over 60 papers. Today, his main interests are in the application of Systems Theory, Frameworks and the Soft Systems Methodology.