PREPRINT OF AN ARTICLE IN: ICCA JOURNAL 21(4), PP. 228–244, DEC. 1998.

DarkThought Goes Deep

Ernst A. Heinz*
Institute for Program Structures and Data Organization (IPD)
School of Computer Science, University of Karlsruhe
P.O. Box 6980, D-76128 Karlsruhe, F.R. Germany

November 30, 1998

Abstract

This article follows up on the work by Hyatt and Newborn who let Hyatt's chess program CRAFTY "go deep" in 1997. We repeated Hyatt and Newborn's experiment with our own chess program DARKTHOUGHT[1] and obtained similar experimental results as originally reported by Hyatt and Newborn with CRAFTY. Both experiments provide strong empirical evidence for the surprising observation that even at high search depths of 11–14 plies modern chess programs steadily discover new best moves in still 16% of all searches on average.

Moreover, the experiments do not reveal any conclusive trend towards fewer new best moves at search depths beyond 14 plies. Hence, the available experimental results do not really fuel the intuitive notion that such "changes of mind" taper off continuously with increasing search depths. If at all, only the behaviour of DARKTHOUGHT with a drop to 13.7% new best moves on average in iteration #14 hinted at decreasing "changes of mind" for search depths of $\geq$ 15 plies.

Additionally gathered data about the 14-ply searches of CRAFTY and DARKTHOUGHT allowed us to study the behaviour of both programs in greater detail. This led to the astonishing finding that, regardless of the actual search depth, sizable 30%–50% of all new best moves on average represented "fresh ideas" which the programs never deemed best before. The finding adds support to Newborn's hypothesis about the playing strength of chess programs as initially formulated in 1985. Furthermore, the new data educated us about continuing odd/even instabilities of modern chess programs. These instabilities decreased notably only at high search depths of 9–14 plies in positions with reduced material as mostly occur in endgames and late middle-games.

*Email = <heinze@ira.uka.de>, WWW = <http://wwwipd.ira.uka.de/~heinze/>
[1]WWW = <http://wwwipd.ira.uka.de/Tichy/DarkThought/>

1 Introduction

Deep searches with far look-ahead continue to fascinate the whole computer-chess community because they promise to achieve ever stronger play for all decent programs. Thompson pioneered the scientific investigation of the general relation between search depth and playing strength in chess programs by his famous self-play experiments with the then reigning World Computer Chess Champion BELLE in the early 1980s [7, 8, 35]. His experiments led to the surprising result that the playing strength of BELLE increased almost linearly with search depth by roughly 200 rating points per ply for full-width searches of fixed depths ranging from 3–9 plies. Several other researchers later confirmed Thompson's findings by self-play experiments with their own chess programs HITECH, PHOENIX, THE TURK, and ZUGZWANG [5, 18, 22]. In Figure 1 of their article [18] Junghanns et al. showed that in all these cases the winning percentages of the program versions searching one ply deeper than their direct siblings remained range-bound between 70%–80% with no clearly visible, let alone statistically significant, average downward trends at the end of the 9-ply data curves. We thoroughly discuss the above mentioned self-play research and further related works in Section 2.

In 1985 Newborn [24] introduced a technique different from self-play in order to study the general relation of search depth and playing strength in chess programs. The rationale of Newborn's novel approach sprang from the assumption that new best moves as discovered by chess programs at higher search depths ought to represent better choices than the best moves preferred at shallower depths. To this end, Newborn tracked the according behaviour of BELLE for searches to fixed depths of 11 plies on a set of 447 test positions from real games. The close correlation of his data with Thompson's earlier self-play results made Newborn propose an interesting hypothesis concerning the playing strength of chess programs that we elaborate on in Section 3.

Unfortunately, Newborn's new experimental methodology of observing the behaviour of deep searches did not gain much popularity. Nobody followed his initial example until Junghanns et al. let PHOENIX and THE TURK search roughly 1,000 positions from self-play games to fixed depths of 9 plies while recording all the new best moves beside other information [18]. In 1997 Hyatt and Newborn himself then conducted another behavioural experiment with Hyatt's chess program CRAFTY searching 347 new test positions to a fixed depth of 14 plies each [17]. This experiment revealed the astonishing fact that the rate of new best moves as chosen by CRAFTY at high search depths of 9–14 plies remained quite steady around 15%–17% on average and hardly decreased anymore. In order to check whether this exceptional behaviour may actually hold for modern chess programs in general or if it was only an artifact of the specific CRAFTY implementation as used by Hyatt and Newborn in 1997, we repeated their experiment with our own chess program DARKTHOUGHT [16].

DARKTHOUGHT is a fast yet sophisticated alpha-beta searcher using PVS/NEGASCOUT [6, 27] with state-of-the-art enhancements like extended and normal futility pruning [15, 29, 36], internal iterative deepening [2, 31], dynamic move ordering (history+killer heuristic) [1, 11, 28, 30, 32], recursive null-move pruning [4, 9, 12], selective extensions [2, 3], interior-node recognizers [14], and an extended transposition table [23, 32]. DARKTHOUGHT routinely reaches 200K nodes per second (nps) and search depths of 11–13 plies in normal middle-game positions at tournament time-controls on a 500MHz DEC Alpha-21164a PC164 workstation. The search speed often doubles in endgame positions and peaks at 750K nps on the aforementioned hardware. Although CRAFTY and DARKTHOUGHT share many basic design principles, they also feature substantial differences regarding such fundamental issues as node expansion, position evaluation, and search strategy. In particular, DARKTHOUGHT seems to be much more selective than CRAFTY in the full-width part of its search (e.g. extended futility pruning). Besides, it relies on lazy alpha-bounding for the top-level search at the root node [16] and thus delays the full resolution of new best moves as long as possible (preferrably to the next iteration). Last but not least, DARKTHOUGHT does not employ fractional extensions which are among the specialities of CRAFTY. Overall we deem it fair to say that both programs sufficiently resemble each other and at the same time differ enough from one another in order to make comparative evaluations of their search depths meaningful.

Hyatt and Newborn relied on more than 20 volunteers to execute the 14-ply searches for the 347 test positions on heterogeneous PC-class machines at speeds of 25K–100K nps. Despite this highly distributed setting, their whole experiment still took about three weeks to complete. We succeeded in repeating the whole experiment in only four and a half days with DARKTHOUGHT running at an average speed of 250K nps each on 8x 500MHz DEC Alpha-21164a PC164 workstations. During the searches we let DARKTHOUGHT record more information about the new best moves at every iteration than what Hyatt and Newborn originally studied for their experiment in 1997. With the help of an automated Perl script we also derived the equivalent additional data for CRAFTY from Hyatt and Newborn's publicly available result file. We present a detailed discussion of our new experimental results for DARKTHOUGHT and CRAFTY in Section 5.

2 Relating Computing Power and Search Depth to the Strength of Chess Programs

To the best of our knowledge, Gillogly and Newborn in 1978 independently reported the earliest attempts at modeling the relationship between the playing strength of chess programs on one hand and the available computing power or search depth on the other. Gillogly introduced his so-called "technology curve" [10] that plotted the playing strength against what he called "machine power" on a logarithmic scale. Newborn related the numbers of nodes as searched by different chess programs in three minutes (average time per move in tournament games) to the playing strengths of the very same programs as derived from their actual performances in tournaments [25, 26]. Later on, Levy [21] and Levy and Newborn [20] refined Newborn's initial scheme by contrasting the highest rated tournament performances of the best chess programs with the years of their achievement. All these comparisons inevitably led to speculative extrapolations of the graphs which Levy characterized exactly to the point as the "meta-science of prediction in computer chess" in his latest article [19] about the subject in 1997.

Self-play matches as pioneered by Thompson with his chess machine BELLE in 1982 [35] represent a more rigorous method of investigating the relationship between available computing power or search depth and the playing strength of chess programs. A great advantage of such matches is that the encountered winning percentages quantify the differences in playing strength of the various participating versions of the same program. Despite lingering doubts and questions regarding the magnitude of self-play rating differences [5], many researchers felt and still feel that self-play is the best of the available approaches in order to assess the expected "diminishing returns" of additional search in chess. In the context of self-play matches the presence of diminishing returns should lead to notably reduced winning percentages of the deeper or longer searching program versions with the progression towards higher search depths and search times.

Up to now nobody proved the existence of diminishing returns for self-play in computer chess by means of statistically significant experiments. For matches of only 20 games, even high winning percentages of 70%–75% barely suffice to decide with good confidence that the victor is indeed stronger than his opponent in general. Consequently, such small sample sizes do not allow for any confident quantification of the absolute rating differences between the opponents. In order to confidently assess rating differences of 70 points we need about 500 games per match and for differences of 40–50 rating points the number of necessary games per match increases to 1000. Hence, we completely agree with Mysliwietz [22] who already criticized the statistical uncertainty of many famous self-play experiments in computer chess back in 1994. Based on these explanations it is now easy to understand why we label many experimental results as "not statistically significant" while describing all prominent self-play experiments in computer chess below.

1982: Thompson [35]. Thompson's pioneering experiment featured 100 self-play games with matches of 20 games each between versions of BELLE differing by exactly one ply in lookahead for fixed search depths of 3–8 plies. The gain in playing strength averaged at 246 rating points per ply of search. The experiment showed no diminishing returns at any search depth but its results are not statistically significant.

1983: Condon and Thompson [7]. In the second experiment, Condon and Thompson let BELLE self-play 300 games in round-robin style with matches of 20 games each between all program versions for fixed search depths of 4–9 plies. The gain in playing strength averaged at 217 rating points per ply of search. Now the observed ratings hinted at limited diminishing returns from a fixed search depth of 6 plies onwards. Yet the results of the experiment are not statistically significant.

1988: Szabo and Szabo [33]. The two Szabos determined the technology curve of their chess program TECHMATE that self-played 6882 games on two Atari ST computers linked together via their MIDI ports. The number of games per match between longer and shorter searching versions of the program varied strongly from a minimum of 32 to a maximum of 1367. The gain in playing strength averaged at 156 rating points per doubling of available search time (computing power). The experimental data indicated slight diminishing returns at longer search times. Several results from the experiment are statistically significant. Unfortunately, however, this does not hold for the results of the most interesting program versions with really long search times. They simply did not play enough games to draw reliable conclusions.

1990: Berliner et al. [5]. The HITECH team made their chess machine self-play 1056 games in a round-robin setting with matches of 16 games each between all program versions of HITECH and LOTECH (a variant of HITECH scaled down knowledge-wise) for fixed search depths of 4–9 plies. The gain in playing strength averaged at 195 rating points per ply of search for HITECH and at 232 rating points per ply for LOTECH. The overall ratings showed signs of limited diminishing returns starting at a fixed search depth of 6 plies. But there was no clear trend of diminishing returns at higher search depths and the experimental results are not statistically significant.

1994: Mysliwietz [22]. In his own experiment, Mysliwietz let the parallel chess program ZUGZWANG self-play 450 games with 50 games per match between program versions that differed roughly by a factor of 2 in search speed due to varying numbers of allotted processors. The gain in playing strength averaged at 109 rating points per doubling of search speed for 9 successive doubling steps. The observed ratings do not exhibit any diminishing returns at all. Most of the match results from this experiment are statistically significant in the sense that they allow for the discrimination of stronger and weaker opponents with 95% confidence. The true general rating gain for the version of ZUGZWANG used in the experiment lay in the estimated range of 76–143 points per ply of search with 95% confidence. Thence, Mysliwietz's experiment does neither support nor falsify the hypothesis of diminishing returns for self-play of ZUGZWANG with good confidence.

1997: Junghanns et al. [18]. The self-play experiment with Junghanns' chess program THE TURK featured 480 games with matches of 80 games each between program versions differing by exactly one ply in lookahead for fixed search depths of 3–9 plies. The gain in playing strength averaged around 200 rating points per ply of search.[2] The winning percentages of the deeper searching versions of THE TURK actually increased steadily from fixed search depths of 6 plies onwards, thus even hinting at additional gains in returns for higher search depths rather than diminishing ones. The match results from this experiment are statistically significant in the sense that they allow for the discrimination of stronger and weaker opponents with 95% confidence. As in the case of Mysliwietz's experiment, however, the resulting data of the THE TURK neither supports nor falsifies the hypothesis of diminishing returns for self-play of this program with good confidence.

In their article Junghanns et al. [18] then continued to look for diminishing returns by means of other metrics than self-play. They finally claimed

[2] We estimated this number because the original text only gives the winning percentages but not the absolute rating differences.

to have found empirical evidence in this respect. According to their explanations the low search quality of chess programs (i.e. their high error probability) and the abnormally large lengths of self-play games inadvertently hide the doubtlessly existent diminishing returns in computer chess. Although we greatly appreciate Junghanns et al.'s new trial aimed at the better understanding of diminishing returns in computer chess, we are not convinced that their claims hold when subjected to rigorous methodological and statistical testing. In our opinion the quest for indisputable and statistically significant demonstrations of diminishing returns for additional search in computer chess still remains to be concluded.

3 Newborn's Hypothesis Concerning the Playing Strength of Chess Programs

In 1985 Newborn discovered that the self-play results of BELLE correlated closely with the rates of new best moves as chosen by it during searches to fixed depths of 11 plies on a set of 447 test positions from real games. Moreover, the behaviour of TECH III [34] during searches to fixed depths of 11 plies for the 300 well-known "Win At Chess" test positions resembled the behaviour of BELLE during its equally deep searches. Based on these two observations Newborn [24] formulated an interesting hypothesis concerning the playing strength of chess programs in general. His hypothesis relates the increases in playing strength for deeper searches to the rates of new best moves as chosen at consecutive search depths (deeper versus shallower). Specifically, let $BC(k)$ denote the rate or number of new best moves as observed in iteration $\#k$ and let $RI(k)$ denote the actual increase in playing strength for searches of depth k in comparison with searches of depth $k-1$.

Newborn's Hypothesis. Suppose values for $BC(k)$, $BC(k+1)$, and $RI(k)$ have been determined over a wide sample of game positions. When going from searches of depth k to searches of depth $k+1$, the rating of the program in question can then be expected to improve by

$$RI(k+1) = RI(k) * \frac{BC(k+1)}{BC(k)}\,.$$

The hypothesis intuitively relies on the straightforward yet very important assumption that the new best moves as determined by deeper searches are of higher quality than the existing best moves which they displace. If this fundamental rationale is not fulfilled, the hypothesis does not make much sense at all. Please imagine a chess program that simply switches back and forth between a few good moves all the time. Such behaviour does surely not increase the playing strength of the program at any search depth. Therefore, the discovery of "fresh ideas" (e.g. new best moves which the program never deemed best before) looks like a much better and more meaningful indicator of increases in playing strength than just the change of the preferred choices. Surprisingly enough, nobody seems to have investigated this aspect of Newborn's hypothesis prior to us. In Section 5.2 we present empirical evidence that supports the validity of the hypothesis in this respect. According to the experimental results of both CRAFTY and DARKTHOUGHT sizable 30%–50% of all new best moves as chosen by modern chess programs represent "fresh ideas" on average.

Newborn's hypothesis neatly fitted the results of the self-play experiments and the behavioural experiments to fixed search depths of 11 plies as available for BELLE in 1985. To the best of our knowledge, however, no researcher ever tested the validity of the hypothesis for higher search depths than 11 plies nor for other chess programs than BELLE. Nevertheless, in 1997 Hyatt and Newborn [17] applied the above formula of expected rating increase for an additional ply of search to their experimental results of CRAFTY in order to calculate "extrapolated" ratings of the program at search depths of 6–14 plies. They seeded their calculations with the according ratings of BELLE from 1985 as given values for the 4-ply and 5-ply searches. Hyatt and Newborn obviously interpreted the similar search behaviours of BELLE and CRAFTY on different sets of test positions as sufficient reason to assume the validity of Newborn's hypothesis for CRAFTY as well.

We object to Hyatt and Newborn's "extrapolation" because in our opinion neither the application of the hypothesis itself nor the repeatedly recursive style of its application were appropriate in this case. First and foremost, there exist no published results of self-play experiments for CRAFTY at any fixed search depths such that not even a single value of $RI(k)$ could be determined in compliance with the prescriptions of the hypothesis. Instead, Hyatt and Newborn rather calculated them too by repeatedly applying the recursive formula of expected rating increase to each additional ply of search. This essentially means that Hyatt and Newborn employed the hypothesis in order to derive assumedly independent preconditions of itself. These fabricated preconditions then served as the basis for the next true application of the hypothesis yielding the calculation of an absolute rating and so forth. Thence, the ratings of CRAFTY as postulated by Hyatt and Newborn in Figure 7 of their article (e.g. 2601 points at a fixed search depth of 11 plies and 2983 points at a fixed search depth of 14 plies) lack real substance.

4 Corrected Test Positions

As offered by Hyatt and Newborn in 1997, we obtained their 347 original test positions from <ftp://ftp.cis.uab.edu/pub/hyatt/plytest/positions.gz>. To our utter suprise we discovered that the original test positions still contained 4 doubles and 4 other errors in their FEN specifications although they were already used in Hyatt and Newborn's experiment. We cross-checked our discovery by inspecting Hyatt and Newborn's publicly available result file from <ftp://ftp.cis.uab.edu/pub/hyatt/plytest/results.gz>. There, the suspected errors and doubles were indeed visible in the output of CRAFTY (wrong castling rights in position #34, Pawn on 8th rank in position #90, and two black Kings in position #255). We removed the doubles and corrected the errors as described below. Overall, we were left with 343 corrected test positions to conduct our own experiment with.[3]

- Doubles occurred in the original positions #1 and #16, #44 and #47, #115 and #279, and #240 and #247 which are pair-wise identical.
- The FEN specification of position #34 mis-spelled the castling rights as `Kqkq` instead of the correct `KQkq`.
- The FEN specification of position #90 placed a non-existent Pawn on the 8th rank by reading `pr4k1/` instead of the correct `1r4k1/`.
- The FEN specification of position #255 placed a second black King on the 1st rank by reading `/1r3k2` instead of the correct `/1r3K2`.
- The move information associated with position #61 suggested to play the impossible `Rxf7` instead of the correct `Ng5`.

5 Experimental Results

Before the start of the experiment we decided to broaden our behavioural studies of deep searches with respect to Hyatt and Newborn's original setup. Because of these extended objectives we let DARKTHOUGHT record more detailed information than just the search depth at which the best move changed as compared with the previous iteration. While searching each of the 343 corrected test positions to a fixed depth of 14 plies, DARKTHOUGHT also determined whether the new best move after the current iteration was "fresh" in the sense that it never constituted the best move at the end of any

[3]Hyatt and Newborn will most probably correct their publicly available files, too.

former iteration. Moreover, DarkThought tracked all the instances where the new best move after the current iteration was the same as the best move of two or three iterations ago. In these cases there are typically several good moves with nearly equal scores that compete for the pole position and cause frequent changes of the best moves in consecutive iterations. The well-known and wide-spread instability of searches at odd and even depths belongs to this behavioural category with the new best move after the current iteration being the same as the best move of two iterations ago.

In order to formalize the above explanations of what DarkThought recorded, we let $B(i)$ denote the best move after iteration i. The columns "Best Change", "Fresh Best", "(I – 2) Best", and "(I – 3) Best" of the forthcoming tables (Table 2 – Table 9) list the numbers of positions for which the following conditions held after the search completed iteration i.

- **Best Change:** $B(i) \neq B(i-1)$.
- **Fresh Best:** $B(i) \neq B(j)$ for all $j < i$.
- **(I – 2) Best:** $B(i) = B(i-2)$ & $B(i) \neq B(i-1)$.
- **(I – 3) Best:** $B(i) = B(i-3)$ & $B(i) \neq B(i-2)$ & $B(i) \neq B(i-1)$.

We also derived these four relations for Crafty by applying a short Perl script to Hyatt and Newborn's publicly available result file (see Section 4). The script automatically computed the complete result data of Crafty as listed in the tables of this section. The data includes all the additional information for Crafty that equals the special recordings of DarkThought. Therefore and because we eliminated the doubles from the original set of test positions, we could not reuse the experimental results as originally published by Hyatt and Newborn in 1997. Furthermore, Hyatt and Newborn's numbers differ from those that our Perl script produces for Crafty from Hyatt and Newborn's original result file which spans the full set of 347 test positions (see Appendix C). Hence, we consistently refer to our own numbers throughout the whole section.

The experimental results as published by other researchers correspond exactly to our "Best Change" columns. Up to now, nobody except Junghanns et al. [18] ever mentioned any further characteristics of the new best moves. In this respect we advance the state-of-the-art by studying some interesting relations between all new best moves ("Fresh Best") and sequences of two or three successive new best moves ("I – 2 / I – 3 Best") that occur during single searches of modern chess programs. The remainder of this section presents the collected data and discusses our interpretation of the raw numbers in detail. For a description of the exact experimental setup of DarkThought please see Appendix A.

5.1 "Best Change" Rates for All Test Positions

Search Depth	Belle 1985	(Standard) (Error)	Crafty 1997	(Standard) (Error)	DarkThought 1998	(Standard) (Error)
2	– –	– –	38.78%	(2.63%)	35.28%	(2.58%)
3	– –	– –	36.73%	(2.60%)	39.65%	(2.64%)
4	33.10%	(2.23%)	30.61%	(2.49%)	31.78%	(2.51%)
5	33.10%	(2.23%)	30.32%	(2.48%)	29.45%	(2.46%)
6	27.70%	(2.12%)	27.41%	(2.41%)	24.49%	(2.32%)
7	29.50%	(2.16%)	24.49%	(2.32%)	21.28%	(2.21%)
8	26.00%	(2.07%)	22.45%	(2.25%)	25.07%	(2.34%)
9	22.60%	(1.98%)	**18.37%**	(2.09%)	21.57%	(2.22%)
10	**17.70%**	(1.81%)	17.20%	(2.04%)	24.20%	(2.31%)
11	18.10%	(1.82%)	16.62%	(2.01%)	**17.49%**	(2.05%)
12	– –	– –	16.91%	(2.02%)	15.45%	(1.95%)
13	– –	– –	14.58%	(1.91%)	16.62%	(2.01%)
14	– –	– –	15.45%	(1.95%)	**13.70%**	(1.86%)

Table 1: "Best Change" Rates of Belle, Crafty, and DarkThought.

Table 1 summarizes the "Best Change" rates $BC(i)$ and their estimated standard deviations = standard errors $s(i) = \sqrt{BC(i) * (1 - BC(i))/343}$ as observed in our experiment for all 343 corrected test positiosn at search depths of 2–14 plies. These percentages of DarkThought closely resemble the according numbers of Crafty from 1997 for the same set of positions and search depths as well as the numbers of Belle from 1985 for a different set of 447 test positions and search depths of 4–11 plies [24]. For the convenience of the reader and in order to make our subsequent discussions more transparent, we also include the numbers of Belle and Crafty in Table 1 showing them side-by-side with our own new data of DarkThought.

The table illustrates that Belle, Crafty, and DarkThought feature very similar "Best Change" behaviours on average. This is quite surprising if you consider the substantial differences of the three programs regarding such fundamental issues as node expansion, position evaluation, and search strategy. The experimental results of DarkThought support the pioneering findings of Hyatt and Newborn at high search depths of 12–14 plies in particular. For these search depths the "Best Change" rates of both Crafty and DarkThought stayed range-bound around 16%. As a tentative conclusion we conjecture that the three columns of Table 1 taken together provide convincing empirical evidence that the very gradual decreases of the "Best Change" rates at high search depths are not only artifacts of specific implementations but rather represent an actually general phenomenon of chess programs which rely on depth-first alpha-beta search with iterative deepening. Despite the overall similarities, however, two numbers of DarkThought roused our attention because they differ notably from those of Belle and Crafty.

Drop below 20%. The "Best Change" rates of both Belle and Crafty dropped at least one iteration earlier to 17%–18% than that of DarkThought (see numbers marked by ** in Table 1) which stayed well above 20% until iteration #10 inclusively. We attribute the more unstable behaviour of DarkThought to the increased selectivity of its search as compared with the two other programs. While the standard errors still leave some room for doubting the statistical significance of the drops below 20%, Appendix B nullifies the corresponding concerns by deriving 80%-confident and 90%-confident bounds on the "Best Change" probabilities of Belle, Crafty, and DarkThought.

Iteration #14. The "Best Change" rates of Crafty remained surprisingly constant at roughly 15%–17% from iteration #9 onwards. DarkThought only behaved like this from iteration #11 to iteration #13 and then recorded another drop of its "Best Change" rate to 13.7% for the final iteration #14 (see number marked by $_{**}$ in Table 1). This constitutes the first experimental result reported so far which hints at the validity of the intuitive notion that the average "Best Change" rates should taper off even further at search depths beyond 14 plies. The experimental results of Crafty do not really support this notion because the "Best Change" rate of Crafty does not decrease but rather increases again for iteration #14. Unfortunately, it remains totally unclear whether the special behaviour of DarkThought signals a consistent trend towards lower "Best Change" rates at higher search depths than 14 plies or if it is just a fluctuation at the end of our data curve. The statistical calculations of Appendix B do not suffice to discriminate the outstanding data point because the 80%-confident and the 90%-confident upper bounds on the "Best Change" probability of DarkThought in iteration #14 equal 15.34% and 16.26% respectively. Thence, new experiments with search depths of at least 16 plies are needed to resolve this interesting question.

5.2 Experimental Results for All Test Positions

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	35.28%	(121)	100.00%	(121)	0.00%	(0)	0.00%	(0)
3	39.65%	(136)	85.29%	(116)	14.71%	(20)	0.00%	(0)
4	31.78%	(109)	55.05%	(60)	31.19%	(34)	13.76%	(15)
5	29.45%	(101)	56.44%	(57)	24.75%	(25)	10.89%	(11)
6	24.49%	(84)	65.48%	(55)	19.05%	(16)	5.95%	(5)
7	21.28%	(73)	49.32%	(36)	28.77%	(21)	10.96%	(8)
8	25.07%	(86)	50.00%	(43)	24.42%	(21)	4.65%	(4)
9	21.57%	(74)	40.54%	(30)	28.38%	(21)	13.51%	(10)
10	24.20%	(83)	37.35%	(31)	34.94%	(29)	8.43%	(7)
11	17.49%	(60)	31.67%	(19)	36.67%	(22)	10.00%	(6)
12	15.45%	(53)	45.28%	(24)	20.76%	(11)	9.43%	(5)
13	16.62%	(57)	42.11%	(24)	28.07%	(16)	10.53%	(6)
14	13.70%	(47)	34.04%	(16)	25.53%	(12)	12.77%	(6)

Table 2: Results of DarkThought for All 343 Corrected Test Positions.

Table 2 summarizes the complete experimental results of DarkThought for all 343 corrected test positions. Table 3 shows the corresponding numbers of Crafty as automatically computed by our Perl script from Hyatt and Newborn's publicly available result file. The three rightmost columns of the tables list the novel statistics of our additionally gathered data. Their percentages relate to the absolute numbers of the "Best Change" column. As already expected beforehand, our novel statistics reveal some very interesting general features of the new best moves at every iteration.

Fresh Best. In contrast to what we and probably many others suspected, the rates of fresh best moves (relative to all new best moves) of both Crafty and DarkThought did not steadily decrease from one iteration to the next – even not at high search depths of 9–14 plies. Instead, the "Fresh Best" rates wavered directionless between 30%–50% from iteration #7 onwards. This finding lends support to the validity of Newborn's hypothesis about the playing strength of chess programs (see Section 3) because

$$\frac{\#\text{“Best Change”}(i)}{\#\text{“Best Change”}(i-1)} \approx \frac{\#\text{“Fresh Best”}(i)}{\#\text{“Fresh Best”}(i-1)} \quad \text{holds for } i \geq 8.$$

The surprising approximation means that the discovery of fresh best moves remains substantial even at high search depths of up to 14 plies

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	38.78%	(133)	100.00%	(133)	0.00%	(0)	0.00%	(0)
3	36.73%	(126)	71.43%	(90)	28.57%	(36)	0.00%	(0)
4	30.61%	(105)	65.71%	(69)	25.71%	(27)	8.57%	(9)
5	30.32%	(104)	59.62%	(62)	30.77%	(32)	6.73%	(7)
6	27.41%	(94)	56.38%	(53)	24.47%	(23)	8.51%	(8)
7	24.49%	(84)	47.62%	(40)	30.95%	(26)	7.14%	(6)
8	22.45%	(77)	37.66%	(29)	31.17%	(24)	11.69%	(9)
9	18.37%	(63)	30.16%	(19)	38.10%	(24)	4.76%	(3)
10	17.20%	(59)	40.68%	(24)	32.20%	(19)	5.08%	(3)
11	16.62%	(57)	52.63%	(30)	24.56%	(14)	5.26%	(3)
12	16.91%	(58)	41.38%	(24)	24.14%	(14)	10.34%	(6)
13	14.58%	(50)	32.00%	(16)	22.00%	(11)	14.00%	(7)
14	15.45%	(53)	39.62%	(21)	32.08%	(17)	5.66%	(3)

Table 3: Results of CRAFTY for All 343 Test Positions.

and decreases as gradually on average as the discovery of any new best moves (see Section 5.1). Given the strong empirical evidence from the experimental results of both CRAFTY and DARKTHOUGHT, we expect the approximation to be valid for modern chess programs in general.

(I – 2) Best. The numbers of this column show that CRAFTY and DARKTHOUGHT suffered from instable odd/even behaviour in 25%–35% of all "Best Change" searches regardless of their nominal depth. We deem it quite remarkable that the rates of odd/even instability (relative to all new best moves) wavered in such a narrow range starting with iteration #3. Overall, the average probabilities of odd/even instability during any "Best Change" search amounted to 26.5% for CRAFTY and 24.4% for DARKTHOUGHT. The experimental results of both programs therefore strongly suggest that modern chess programs feature odd/even instabilities for 25% of all "Best Change" searches in general. Last but not least, we like to mention that the sum of "Fresh Best" moves and "(I – 2) Best" moves equalled about 65%–75% of all new best moves for both CRAFTY and DARKTHOUGHT from iteration #8 onwards (sole exception: CRAFTY at iteration #13).

(I – 3) Best. There are hardly any noteworthy facts to report for the last columns of Table 2 and Table 3. The rates of "(I – 3) Best" moves (relative to all new best moves) wavered more radically than both the "Fresh Best" rates and the "(I – 2) Best" rates. We were surprised that the "(I – 3) Best" rates averaged at 10% for both CRAFTY and DARKTHOUGHT which was well above our expectations.

5.3 Experimental Results for the Opening Positions

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	37.74%	(40)	100.00%	(40)	0.00%	(0)	0.00%	(0)
3	50.94%	(54)	83.33%	(45)	16.67%	(9)	0.00%	(0)
4	39.62%	(42)	54.76%	(23)	35.71%	(15)	9.52%	(4)
5	33.96%	(36)	44.44%	(16)	38.89%	(14)	0.00%	(0)
6	25.47%	(27)	55.56%	(15)	25.93%	(7)	3.70%	(1)
7	16.98%	(18)	44.44%	(8)	27.78%	(5)	22.22%	(4)
8	25.47%	(27)	59.26%	(16)	11.11%	(3)	7.41%	(2)
9	25.47%	(27)	37.04%	(10)	29.63%	(8)	11.11%	(3)
10	29.25%	(31)	25.81%	(8)	41.94%	(13)	9.68%	(3)
11	18.87%	(20)	15.00%	(3)	55.00%	(11)	15.00%	(3)
12	17.93%	(19)	42.11%	(8)	26.32%	(5)	15.79%	(3)
13	16.04%	(17)	41.18%	(7)	35.29%	(6)	5.88%	(1)
14	13.21%	(14)	35.71%	(5)	14.29%	(2)	14.29%	(2)

Table 4: Results of DARKTHOUGHT for the 106 Opening Positions.

The 343 corrected test positions contain a subset of 106 opening positions that occurred at the 7th move (51 positions) and the 14th move (55 positions) of real chess games. Table 4 presents the experimental results of DARKTHOUGHT for these 106 opening positions alone. Table 5 lists the corresponding numbers of CRAFTY as automatically computed by our Perl script from Hyatt and Newborn's publicly available result file. The numbers exhibit the following unusual yet clearly sporadic fluctuations.

- DARKTHOUGHT – "Best Change" at iteration #7, "Fresh Best" at iteration #11, "(I – 2) Best" at iteration #8, and "(I – 2) Best" iteration #11 and iteration #14.
- CRAFTY – "Fresh Best" at iter. #9 and "(I – 2) Best" at iteration #6.

We attribute these fluctuations to the reduced size of the test subset. Other than that the results for the opening positions alone closely resemble those for the whole test set which Section 5.1 and Section 5.2 already discussed in great detail. The analyses of the overall results mostly apply to the results for the opening subset as well. Only the "Best Change" rates of

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	34.91%	(37)	100.00%	(37)	0.00%	(0)	0.00%	(0)
3	49.06%	(52)	65.38%	(34)	34.62%	(18)	0.00%	(0)
4	32.08%	(34)	55.88%	(19)	38.24%	(13)	5.88%	(2)
5	33.02%	(35)	60.00%	(21)	31.43%	(11)	8.57%	(3)
6	31.13%	(33)	60.61%	(20)	15.15%	(5)	12.12%	(4)
7	30.19%	(32)	43.75%	(14)	34.38%	(11)	9.38%	(3)
8	25.47%	(27)	29.63%	(8)	44.44%	(12)	11.11%	(3)
9	22.64%	(24)	12.50%	(3)	41.67%	(10)	4.17%	(1)
10	19.81%	(21)	52.38%	(11)	28.57%	(6)	4.76%	(1)
11	19.81%	(21)	38.10%	(8)	23.81%	(5)	14.29%	(3)
12	22.64%	(24)	45.83%	(11)	33.33%	(8)	0.00%	(0)
13	16.98%	(18)	27.78%	(5)	22.22%	(4)	22.22%	(4)
14	19.81%	(21)	33.33%	(7)	38.10%	(8)	4.76%	(1)

Table 5: Results of Crafty for the 106 Opening Positions.

Crafty for the opening positions do not really fit into the general picture because they stayed around 20% up to iteration #14 inclusively. In contrast to the overall results, DarkThought executed much stabler searches in the opening positions than Crafty from iteration #11 onwards. This is probably due to a more aggressive evaluation of development and tempi of Crafty in the opening phase of the game. The special behaviour of DarkThought that drops its "Best Change" rate to nearly 13% in iteration #14 is also visible for the opening subset alone.

5.4 Experimental Results for the Middle-Game Positions

The 343 corrected test positions contain a subset of 107 middle-game positions that occurred at the 21st move (55 positions) and the 28th move (52 positions) of real chess games. Table 6 presents the experimental results of DarkThought for these 107 middle-game positions alone. Table 7 lists the corresponding numbers of Crafty as automatically computed by our Perl script from Hyatt and Newborn's publicly available result file. The numbers exhibit the following unusual yet clearly sporadic fluctuations.

- DarkThought – "Best Change" at iteration #12 and iteration #13, "Fresh Best" at iteration #6 and iteration #12, "(I – 2) Best" at iteration #6, and "(I – 2) Best" at iteration #12 and iteration #14.
- Crafty – "Fresh Best" at iteration #11 and iteration #14.

Search Depth	Best Change	(#)	Fresh Best	(#)	(I − 2) Best	(#)	(I − 3) Best	(#)
2	28.97%	(31)	100.00%	(31)	0.00%	(0)	0.00%	(0)
3	40.19%	(43)	86.05%	(37)	13.95%	(6)	0.00%	(0)
4	27.10%	(29)	51.72%	(15)	31.03%	(9)	17.24%	(5)
5	31.78%	(34)	55.88%	(19)	20.59%	(7)	20.59%	(7)
6	20.56%	(22)	81.82%	(18)	18.18%	(4)	9.09%	(2)
7	24.30%	(26)	46.15%	(12)	30.77%	(8)	11.54%	(3)
8	28.97%	(31)	54.84%	(17)	29.03%	(9)	0.00%	(0)
9	23.36%	(25)	32.00%	(8)	36.00%	(9)	16.00%	(4)
10	27.10%	(29)	41.38%	(12)	44.83%	(13)	6.90%	(2)
11	18.69%	(20)	30.00%	(6)	35.00%	(7)	5.00%	(1)
12	12.15%	(13)	53.85%	(7)	15.38%	(2)	7.69%	(1)
13	23.36%	(25)	40.00%	(10)	28.00%	(7)	12.00%	(3)
14	14.95%	(16)	25.00%	(4)	50.00%	(8)	6.25%	(1)

Table 6: Results of DarkThought for the 107 Middle-Game Positions.

We attribute these fluctuations to the reduced size of the test subset as in the case of the opening positions (see Section 5.3). Other than that the results for the middle-game positions alone closely resemble those for the whole test set which Section 5.1 and Section 5.2 already discussed in great detail. The analyses of the overall results mostly apply to the results for the middle-game subset as well. However, the "Fresh Best" rates of Crafty and DarkThought remained larger on average for the middle-game positions alone than for all other positions at high search depths of 9–14 plies. As probably expected by most experts, the real middle-game positions were therefore clearly the hardest for both programs to decide on. Moreover, the behaviour of Crafty for the middle-game subset does not lead to an increase of its "Best Change" rate at the final iteration #14.

5.5 Experimental Results for the Remaining Positions

When removing the 106 opening positions and the 107 middle-game positions from the full set of 343 corrected test positions, you end up with a subset of 130 remaining positions that occurred at the 35th, 42nd, ..., and 70th moves of real chess games. This final subset contains mostly endgame positions complemented by a few ones from late middle-games. Table 8 presents the experimental results of DarkThought for these 130 remaining positions alone. Table 9 lists the corresponding numbers of Crafty as automatically computed by our Perl script from Hyatt and Newborn's

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	35.51%	(38)	100.00%	(38)	0.00%	(0)	0.00%	(0)
3	31.78%	(34)	76.47%	(26)	23.53%	(8)	0.00%	(0)
4	33.64%	(36)	58.33%	(21)	30.56%	(11)	11.11%	(4)
5	35.51%	(38)	52.63%	(20)	34.21%	(13)	7.89%	(3)
6	28.04%	(30)	56.67%	(17)	30.00%	(9)	10.00%	(3)
7	21.50%	(23)	47.83%	(11)	34.78%	(8)	8.70%	(2)
8	21.50%	(23)	43.48%	(10)	26.09%	(6)	4.35%	(1)
9	20.56%	(22)	36.36%	(8)	45.45%	(10)	4.55%	(1)
10	21.50%	(23)	30.43%	(7)	34.78%	(8)	8.70%	(2)
11	18.69%	(20)	60.00%	(12)	30.00%	(6)	0.00%	(0)
12	16.82%	(18)	33.33%	(6)	22.22%	(4)	22.22%	(4)
13	13.08%	(14)	35.71%	(5)	28.57%	(4)	7.14%	(1)
14	13.08%	(14)	50.00%	(7)	28.57%	(4)	7.14%	(1)

Table 7: Results of CRAFTY for the 107 Middle-Game Positions.

publicly available result file. The numbers exhibit the following unusual yet clearly sporadic fluctuations.

- DARKTHOUGHT – "Fresh Best" at iteration #8 and "(I – 2) Best" at iteration #10 and iteration #14.
- CRAFTY – "Fresh Best" at iteration #11, "(I – 2) Best" at iteration #10, and "(I – 2) Best" at iteration #12 and iteration #14.

We attribute these fluctuations to the reduced size of the test subset as in the cases of the opening and the middle-game positions (see Section 5.3 and Section 5.4). Up to iteration #8 inclusively the results for the remaining positions alone closely resemble those for the whole test set which Section 5.1 and Section 5.2 already discussed in great detail. Starting with iteration #9, however, the "Best Change" rates of CRAFTY and DARKTHOUGHT for the remaining subset dropped markedly below that of all other positions and stayed lower on average until iteration #14. Given the strong empirical evidence from the experimental results of both CRAFTY and DARKTHOUGHT, we expect this more stable behaviour at high search depths in late middle-games and endgames to hold for modern chess programs in general.

Thanks to our additionally recorded data we can actually explain the drops of the "Best Change" rates from iteration #9 onwards. In comparison with their preceding siblings for the opening and the middle-game positions Table 8 and Table 9 reveal that the numbers of "(I – 2) Best"

Search Depth	Best Change	(#)	Fresh Best	(#)	(I – 2) Best	(#)	(I – 3) Best	(#)
2	38.46%	(50)	100.00%	(50)	0.00%	(0)	0.00%	(0)
3	30.00%	(39)	87.18%	(34)	12.82%	(5)	0.00%	(0)
4	29.23%	(38)	57.89%	(22)	26.32%	(10)	15.79%	(6)
5	23.85%	(31)	70.97%	(22)	12.90%	(4)	12.90%	(4)
6	26.92%	(35)	62.86%	(22)	14.29%	(5)	5.71%	(2)
7	22.31%	(29)	62.07%	(18)	27.59%	(8)	3.45%	(1)
8	21.54%	(28)	35.71%	(10)	32.14%	(9)	7.14%	(2)
9	16.92%	(22)	54.55%	(12)	18.18%	(4)	13.64%	(3)
10	17.69%	(23)	47.83%	(11)	13.04%	(3)	8.70%	(2)
11	15.38%	(20)	50.00%	(10)	20.00%	(4)	10.00%	(2)
12	16.15%	(21)	42.86%	(9)	19.05%	(4)	4.76%	(1)
13	11.54%	(15)	46.67%	(7)	20.00%	(3)	13.33%	(2)
14	13.08%	(17)	41.18%	(7)	11.76%	(2)	17.65%	(3)

Table 8: Results of DarkThought for the 130 Remaining Positions.

moves for the subset of remaining positions differ substantially from those of the other subsets. Starting with iteration #9, both Crafty and DarkThought produced only three "(I – 2) Best" moves per iteration on average for the remaining subset (50%–100% less than usual) while the overall numbers of "Fresh Best" moves and "(I – 3) Best" moves remained roughly the same as for the other subsets.[4] Therefore, the notable drops of the "Best Change" rates spring almost entirely from the smaller number of "(I – 2) Best" moves. This means that the odd/even instability of both Crafty and DarkThought decreased with less material on the chess board.

6 Conclusion

The repetition of Hyatt and Newborn's experiment with our own chess program DarkThought confirmed their findings regarding the "Best Change" rates of Crafty. Based on the experimental results of both Crafty and DarkThought taken together, we are confident to project steady discoveries of new best moves in at least 16% of all searches on average for modern chess programs even at high search depths of 11–14 plies. Surprisingly enough, the experiments do not provide any conclusive empirical evidence for the intuitive notion that the "Best Change" rates taper off continuously

[4] Of course, the lower numbers of "(I – 2) Best" moves caused the "Fresh Best" rates to increase and the "(I – 2) Best" rates to decrease w.r.t. the overall "Best Change" amounts.

Search Depth	**Best Change**	(#)	**Fresh Best**	(#)	**(I – 2) Best**	(#)	**(I – 3) Best**	(#)
2	44.62%	(58)	100.00%	(58)	0.00%	(0)	0.00%	(0)
3	30.77%	(40)	75.00%	(30)	25.00%	(10)	0.00%	(0)
4	26.92%	(35)	82.86%	(29)	8.57%	(3)	8.57%	(3)
5	23.85%	(31)	67.74%	(21)	25.81%	(8)	3.23%	(1)
6	23.85%	(31)	51.61%	(16)	29.03%	(9)	3.23%	(1)
7	22.31%	(29)	51.72%	(15)	24.14%	(7)	3.45%	(1)
8	20.77%	(27)	40.74%	(11)	22.22%	(6)	14.81%	(4)
9	13.08%	(17)	47.06%	(8)	23.53%	(4)	5.88%	(1)
10	11.54%	(15)	40.00%	(6)	33.33%	(5)	0.00%	(0)
11	12.31%	(16)	62.50%	(10)	18.75%	(3)	0.00%	(0)
12	12.31%	(16)	43.75%	(7)	12.50%	(2)	22.22%	(2)
13	13.85%	(18)	33.33%	(6)	16.67%	(3)	12.50%	(2)
14	13.85%	(18)	38.89%	(7)	27.78%	(5)	5.56%	(1)

Table 9: Results of CRAFTY for the 130 Remaining Positions.

with increasing search depths. They rather remained range-bound within 15%–17% for most of the deep searches as investigated in the experiments. If at all, primarily the behaviour of DARKTHOUGHT with a drop to roughly 13.7% new best moves on average in iteration #14 hinted at decreasing "changes of mind" for search depths of 15 plies and more. Further experiments with markedly higher search depths of at least 16 plies are needed to resolve the interesting question whether this signals a consistent trend towards even lower "Best Change" rates at search depths beyond 14 plies or if it was just a fluctuation at the end of our data curve.

At this point we take the opportunity in order to call on all prospective experimenters to derive or record at least as much data as we did for the experiments with CRAFTY and DARKTHOUGHT.[5] Studying the rates of "Fresh Best" moves and "(I – 2) Best" moves will surely be worth the additional efforts without any doubt. These rates revealed novel traits and insights concerning the search behaviours of both CRAFTY and DARKTHOUGHT. The "Fresh Best" rates led to the astonishing observation that regardless of the actual search depth sizable 30%–50% of all new best moves on average represented "fresh ideas" which the programs never deemed best before. This finding supports the validity of Newborn's hypothesis about the play-

[5]The complete yet totally raw result files of the experiments with both CRAFTY and DARKTHOUGHT as well as our short Perl script for the automatic analysis of the result files of CRAFTY are publicly available in gzip'ed TAR format at URL <ftp://i41s10.ira.uka.de/pub/ernst/deep14.tgz>.

ing strength of chess programs as originally formulated in 1985. The "(I – 2) Best" rates educated us about continuing search instabilities of modern chess programs at odd and even search depths respectively. These odd/even instabilities decreased solely at high search depths of 9–14 plies in positions with reduced material as found mostly in endgames and late middle-games.

Last but not least, we like to mention that the repetition of Hyatt and Newborn's experiment was not only a scientifically rewarding but also a practically useful endeavour. For what could better demonstrate the effective scalability of DarkThought than to "go deep" at fixed search depths of 14 plies on 343 test positions from real games in less than five days?

7 Acknowledgements

The Institute for Program Structures and Data Organization at the University of Karlsruhe (Chair Prof. Tichy) generously provided 8x 500MHz DEC Alpha-21164a PC164 workstations for our exclusive use during a whole week in July 1998. Without these abundant computing resources the execution of our "deep" experiment with DarkThought would have been much harder to undertake. The anonymous reviewers deserve due credits and sincere thanks for their valuable comments and suggestions which greatly helped to improve the initial submission. The first raw version of the Perl script that we employed for the automatic analysis of the experimental results of Crafty stems from one of the reviewers. Finally, we warmly thank Prof. van dan Herik (Editor-in-Chief) and the remaining editorial staff of the ICCA Journal for their superb work and kind cooperation which enabled the inclusion of some late additions and revisions in this article.

A Experimental Setup

- 343 test positions as obtained and corrected (see Section 4) from <ftp://ftp.cis.uab.edu/pub/hyatt/plytest/positions.gz>,
- DarkThought as of July 31, 1998 with 8M transposition-table entries (4M per side), 1M King hash-table entries (512K per side), and 512K Pawn hash-table entries,
- Digital Unix 4.0d program development tools and operating system,
- 500MHz AlphaPC164 workstation (500MHz DEC Alpha-21164a CPU, 8KB/8KB on-chip I/D L1 caches, 96KB unified on-chip L2 cache,

1MB unified off-chip L3 cache, DEC PC164 mainboard, 8x32MB PS/2 SIMMs = 256MB RAM).

B Confident Bounds on the "Best Change" Probabilities

The "Best Change" behaviours of chess programs represent typical count data for a binary-valued random variable in the terms of standard statistics. The count probabilities of binary-valued random variables generally adhere to binomial distributions. For large enough sample sizes n and success counts m with $m > 4$ and $n - m > 4$, however, corresponding normal distributions provide practically sufficient approximations of the awkward to handle binomial distributions. Classic engineering statistics [13] derive the following lower and upper bounds of the success probability $P_{\%}$ for given values of m, n and any desired %-level of confidence as specified by the single-sided percentiles $\lambda_{\%}$ of the $N(0,1)$ normal distribution.

$$P_{\%} \;\geq\; \left(m + \lambda_{\%}^2/2 - \lambda_{\%} * \sqrt{m * (1 - m/n) + \lambda_{\%}^2/4}\right) * (n + \lambda_{\%}^2)^{-1} \qquad (1)$$

$$P_{\%} \;\leq\; \left(m + \lambda_{\%}^2/2 + \lambda_{\%} * \sqrt{m * (1 - m/n) + \lambda_{\%}^2/4}\right) * (n + \lambda_{\%}^2)^{-1} \qquad (2)$$

With the help of these formulas we determined 80%-confident ($\lambda_{\%} = 0.8416$) and 90%-confident ($\lambda_{\%} = 1.2816$) bounds on the "Best Change" probabilities of Belle ($n = 447$), Crafty ($n = 343$), and DarkThought ($n = 343$). For Belle we calculated the success count m from its "Best Change" rates of Table 1. For Crafty and DarkThought we used the absolute "Best Change" numbers of Table 2 and Table 3 as their observed success counts m. The resulting bounds clearly discriminate the drops of the "Best Change" rates below 20% for all three programs with at least 80% confidence (see Table 10: $\trianglerighteq$ = lower bound, $\leq$ = upper bound).

C Experimental Results of Crafty as Published by Hyatt and Newborn in 1997

Table 11 lists the "Best Change" rates as originally published by Hyatt and Newborn in 1997 [17]. Because Hyatt and Newborn did not provide any explicit numbers for the 107 middle-game positions alone, we calculated them as described in Appendix C.1 below. The numbers in parantheses represent

Search Depth	**Belle C = 80%**	**Belle C = 90%**	**Crafty C = 80%**	**Crafty C = 90%**	**DarkTh. C = 80%**	**DarkTh. C = 90%**
2	– –	– –	$\trianglerighteq$ 36.59%	$\trianglerighteq$ 35.46%	$\trianglerighteq$ 33.14%	$\trianglerighteq$ 32.05%
3	– –	– –	$\trianglerighteq$ 34.57%	$\trianglerighteq$ 33.47%	$\trianglerighteq$ 37.45%	$\trianglerighteq$ 36.32%
4	$\trianglerighteq$ 31.26%	$\trianglerighteq$ 30.32%	$\trianglerighteq$ 28.56%	$\trianglerighteq$ 27.52%	$\trianglerighteq$ 29.70%	$\trianglerighteq$ 28.65%
5	$\trianglerighteq$ 31.26%	$\trianglerighteq$ 30.32%	$\trianglerighteq$ 28.27%	$\trianglerighteq$ 27.24%	$\trianglerighteq$ 27.42%	$\trianglerighteq$ 26.40%
6	$\trianglerighteq$ 25.99%	$\trianglerighteq$ 25.11%	$\trianglerighteq$ 25.43%	$\trianglerighteq$ 24.43%	$\trianglerighteq$ 22.59%	$\trianglerighteq$ 21.64%
7	$\trianglerighteq$ 27.75%	$\trianglerighteq$ 26.84%	$\trianglerighteq$ 22.59%	$\trianglerighteq$ 21.64%	$\trianglerighteq$ 19.48%	$\trianglerighteq$ 18.59%
8	$\trianglerighteq$ 24.24%	$\trianglerighteq$ 23.38%	$\trianglerighteq \underline{20.61\%}$	$\trianglerighteq$ 19.70%	$\trianglerighteq$ 23.16%	$\trianglerighteq$ 22.20%
9	$\trianglerighteq \underline{20.97\%}$	$\trianglerighteq$ 20.16%	$\leq$ *20.19%*	$\leq$ *21.20%*	$\trianglerighteq$ 19.76%	$\trianglerighteq$ 18.87%
10	$\leq$ *19.24%*	$\leq$ *20.10%*	$\leq$ *18.98%*	$\leq$ *19.97%*	$\trianglerighteq \underline{22.31\%}$	$\trianglerighteq$ 21.36%
11	$\leq$ *19.70%*	$\leq$ *20.57%*	$\leq$ *18.38%*	$\leq$ *19.35%*	$\leq$ *19.29%*	$\leq$ *20.27%*
12	– –	– –	$\leq$ *18.68%*	$\leq$ *19.66%*	$\leq$ *17.17%*	$\leq$ *18.12%*
13	– –	– –	$\leq$ *16.25%*	$\leq$ *17.19%*	$\leq$ *18.38%*	$\leq$ *19.35%*
14	– –	– –	$\leq$ *17.17%*	$\leq$ *18.12%*	$\leq$ *15.34%*	$\leq$ *16.26%*

Table 10: Confident Bounds on the "Best Change" Probabilities.

the rates that our automatic Perl script computed when we applied it to Hyatt and Newborn's original result file of CRAFTY for the full set of 347 test positions. The slight differences do not affect the overall interpretation of the results and we regard them as constituting some omnipresent imprecisions of measurement. Please note that we debugged our Perl script thoroughly, performed many cross-checks during its execution and really trust its automatically generated output.

C.1 Results of Crafty for the Middle-Game Positions as Derived from Hyatt and Newborn's Publication

Unfortunately, Hyatt and Newborn did not present the experimental results of CRAFTY for the 107 middle-game positions alone in their article. Therefore, we derived the "Best Change" rates of CRAFTY for these positions by ourselves from the available data of the other positions. The absolute number of new best moves as chosen by CRAFTY in the middle-game positions equals the number of new best moves for all test positions minus the new best moves for the opening positions minus the new best moves of the remaining positions. Table 12 shows our according calculations in detail and lists the resulting "Best Change" rates of CRAFTY for the 107 middle-game positions from iteration #2 to iteration #14.

Search Depth	Crafty All	(Auto-)(matic)	Crafty Open	(Auto-)(matic)	Crafty Middle	(Auto-)(matic)	Crafty Rest	(Auto-)(matic)
2	38.90%	(38.90%)	35.50%	(35.45%)	35.50%	(35.51%)	44.60%	(44.62%)
3	37.80%	(36.89%)	49.10%	(49.09%)	33.80%	(31.78%)	31.50%	(30.77%)
4	30.00%	(30.55%)	30.00%	(31.82%)	34.60%	(33.64%)	26.20%	(26.92%)
5	30.50%	(30.55%)	33.60%	(33.64%)	35.50%	(35.51%)	23.80%	(23.85%)
6	27.40%	(27.38%)	30.00%	(30.91%)	28.10%	(28.04%)	24.60%	(23.85%)
7	23.30%	(25.07%)	30.00%	(31.82%)	19.50%	(21.50%)	20.80%	(22.31%)
8	23.30%	(22.48%)	26.40%	(25.45%)	24.10%	(21.50%)	20.00%	(20.77%)
9	17.30%	(18.16%)	21.80%	(21.82%)	19.70%	(20.56%)	11.50%	(13.08%)
10	17.90%	(17.00%)	20.00%	(19.09%)	22.50%	(21.50%)	12.30%	(11.54%)
11	16.70%	(16.43%)	20.00%	(19.09%)	17.70%	(18.69%)	13.10%	(12.31%)
12	17.00%	(17.00%)	23.60%	(22.73%)	15.90%	(16.82%)	12.30%	(12.31%)
13	14.40%	(14.70%)	17.30%	(17.27%)	13.00%	(13.08%)	13.10%	(13.85%)
14	15.30%	(15.56%)	19.10%	(20.00%)	13.20%	(13.08%)	13.80%	(13.85%)

Table 11: "Best Change" Rates as Published by Hyatt and Newborn.

References

[1] Akl, S. G. and Newborn, M. M. (1977). The principal continuation and the killer heuristic. *ACM '77 National Conference*, Proceedings, pp. 466–473, ACM Press.

[2] Anantharaman, T. S. (1991). Extension heuristics. *ICCA Journal*, Vol. 14, No. 2, pp. 47–65.

[3] Beal, D. F. and Smith, M. C. (1995). Quantification of search extension benefits. *ICCA Journal*, Vol. 18, No. 4, pp. 205–218.

[4] Beal, D. F. (1989). Experiments with the null move. *Advances in Computer Chess 5*, D. F. Beal (ed.), pp. 65–79, Elsevier Science, ISBN 0-444-87159-4.

[5] Berliner, H. J. and Goetsch, G. and Campbell, M. S. and Ebeling, C. (1990). Measuring the performance potential of chess programs. *Artificial Intelligence*, Vol. 43, No. 1, pp. 7–21.

[6] Campbell, M. S. and Marsland, T. A. (1983). A comparison of minimax tree search algorithms. *Artificial Intelligence*, Vol. 20, No. 4, pp. 347–367.

[7] Condon, J. H. and Thompson, K. (1983). BELLE. *Chess Skill in Man and Machine*, P. W. Frey (ed.), pp. 82–118, Springer, 2nd ed. 1983, ISBN 0-387-90790-4/3-540-90790-4.

[8] Condon, J. H. and Thompson, K. (1983). BELLE chess hardware. *Advances in Computer Chess 3*, M. R. B. Clarke (ed.), pp. 45–54, Pergamon, ISBN 0-

Search Depth	Calculation of the "Best Change" Rates for the 107 Middle-Game Positions		Best Change
2	$(0.389 * 347 - 0.355 * 110 - 0.446 * 130) \ / \ 107$	=	35.50%
3	$(0.378 * 347 - 0.491 * 110 - 0.315 * 130) \ / \ 107$	=	33.80%
4	$(0.300 * 347 - 0.300 * 110 - 0.262 * 130) \ / \ 107$	=	34.60%
5	$(0.305 * 347 - 0.336 * 110 - 0.238 * 130) \ / \ 107$	=	35.50%
6	$(0.274 * 347 - 0.300 * 110 - 0.246 * 130) \ / \ 107$	=	28.10%
7	$(0.233 * 347 - 0.300 * 110 - 0.208 * 130) \ / \ 107$	=	19.50%
8	$(0.233 * 347 - 0.264 * 110 - 0.200 * 130) \ / \ 107$	=	24.10%
9	$(0.173 * 347 - 0.218 * 110 - 0.115 * 130) \ / \ 107$	=	19.70%
10	$(0.179 * 347 - 0.200 * 110 - 0.123 * 130) \ / \ 107$	=	22.50%
11	$(0.167 * 347 - 0.200 * 110 - 0.131 * 130) \ / \ 107$	=	17.70%
12	$(0.170 * 347 - 0.236 * 110 - 0.123 * 130) \ / \ 107$	=	15.90%
13	$(0.144 * 347 - 0.173 * 110 - 0.131 * 130) \ / \ 107$	=	13.00%
14	$(0.153 * 347 - 0.191 * 110 - 0.138 * 130) \ / \ 107$	=	13.20%

Table 12: Calculated Results of CRAFTY for the 107 Middle-Game Positions.

080-26898-6 [reprinted in *Computer Chess Compendium*, D. N. L. Levy (ed.), pp. 286–292, Springer, 1989, ISBN 0-387-91331-9].

[9] Donninger, C. (1993). Null move and deep search: Selective search heuristics for obtuse chess programs. *ICCA Journal*, Vol. 16, No. 3, pp. 137–143.

[10] Gillogly. J. J. (1978). *Performance Analysis of the Technology Chess Program*. Ph.D. Thesis, Carnegie-Mellon University [printed as Technical Report CMU-CS-78-189, Computer Science Department, Carnegie-Mellon University].

[11] Gillogly, J. J. (1972). The technology chess program. *Artificial Intelligence*, Vol. 3, No. 1–3, pp. 145–163 [reprinted in *Computer Chess Compendium*, D. N. L. Levy (ed.), pp. 67–79, Springer, 1989, ISBN 0-387-91331-9].

[12] Goetsch, G. and Campbell, M. S. (1990). Experiments with the null-move heuristic. *Computers, Chess, and Cognition*, T. A. Marsland and J. Schaeffer (eds.), pp. 159–168, Springer, ISBN 0-387-97415-6/3-540-97415-6.

[13] Heinhold, J. and Gaede, K.-W. (1964). *Ingenieur-Statistik*. Oldenbourg, 3rd ed. 1972, ISBN 3-486-31743-1.

[14] Heinz, E. A. (1998). Efficient interior-node recognition. *ICCA Journal*, Vol. 21, No. 3, pp. 156–167.

[15] Heinz, E. A. (1998). Extended futility pruning. *ICCA Journal*, Vol. 21, No. 2, pp. 75–83.

[16] Heinz, E. A. (1997). How DARKTHOUGHT plays chess. *ICCA Journal*, Vol. 20, No. 3, pp. 166–176.

[17] Hyatt, R. M. and Newborn, M. M. (1997). CRAFTY goes deep. *ICCA Journal*, Vol. 20, No. 2, pp. 79–86.

[18] Junghanns, A. and Schaeffer, J. and Brockington, M. and Björnsson, Y. and Marsland, T. A. (1997). Diminishing returns for additional search in chess. *Advances in Computer Chess 8*, H. J. van den Herik and J. W. H. M. Uiterwijk (eds.), pp. 53–67, University of Maastricht, ISBN 9-062-16234-7.

[19] Levy, D. N. L. (1997). Crystal balls: The meta-science of prediction in computer chess. *ICCA Journal*, Vol. 20, No. 2, pp. 71–78.

[20] Levy, D. N. L. and Newborn, M. M. (1991). *How Computers Play Chess*. Computer Science Press, ISBN 0-716-78121-2/0-716-78239-1.

[21] Levy, D. N. L. (1986). When will brute force programs beat Kasparov? *ICCA Journal*, Vol. 9, No. 2, pp. 81–86.

[22] Mysliwietz, P. (1994). *Konstruktion und Optimierung von Bewertungsfunktionen beim Schach*. Dissertation (Ph.D. Thesis), University of Paderborn.

[23] Nelson, H. L. (1985). Hash tables in CRAY BLITZ. *ICCA Journal*, Vol. 8, No. 1, pp. 3–13.

[24] Newborn, M. M. (1985). A hypothesis concerning the strength of chess programs. *ICCA Journal*, Vol. 8, No. 4, pp. 209–215.

[25] Newborn, M. M. (1979). Recent progress in computer chess. *Advances in Computers*, Vol. 18, No. ?, pp. 59–117 [reprinted in *Computer Games I*, D. N. L. Levy (ed.), pp. 226–324, Springer, ISBN 0-387-96496-4/3-540-96496-4].

[26] Newborn, M. M. (1978). Computer chess: Recent progress and future expectations. *3rd Jerusalem Conference on Information Technology*, Proceedings, J. Moneta (ed.), pp. ?-?, North-Holland, ISBN 0-444-85192-5.

[27] Reinefeld, A. (1983). An improvement to the SCOUT tree-search algorithm. *ICCA Journal*, Vol. 6, No. 4, pp. 4–14.

[28] Schaeffer, J. (1989). The history heuristic and alpha-beta search enhancements in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 11, pp. 1203–1212.

[29] Schaeffer, J. (1986). *Experiments in Search and Knowledge*. Ph.D. Thesis, University of Waterloo [reprinted as Technical Report TR 86–12, Department of Computing Science, University of Alberta].

[30] Schaeffer, J. (1983). The history heuristic. *ICCA Journal*, Vol. 6, No. 3, pp. 16–19.

[31] Scott, J. J. (1969). A chess-playing program. *Machine Intelligence 4*, B. Meltzer and D. Michie (eds.), pp. 255–265, Edinburgh University Press, 1973, ISBN 0-852-24062-7.

[32] Slate, D. J. and Atkin, L. R. (1977). CHESS 4.5 – The Northwestern University chess program. *Chess Skill in Man and Machine*, P. W. Frey (ed.), pp. 82–118, Springer, 2nd ed. 1983, ISBN 0-387-90790-4/3-540-90790-4.

[33] Szabo, A. and Szabo, B. (1988). The technology curve revisited. *ICCA Journal*, Vol. 11, No. 1, pp. 14–20.

[34] Szabo, A. (1984). *Computer-Chess Tactics and Strategy*. M.Sc. Thesis, University of British Columbia.

[35] Thompson, K. (1982). Computer chess strength. *Advances in Computer Chess 3*, M. R. B. Clarke (ed.), pp. 55–56, Pergamon, ISBN 0-080-26898-6.

[36] Ye, C. and Marsland, T. A. (1992). Experiments in forward pruning with limited extensions. *ICCA Journal*, Vol. 15, No. 2, pp. 55–66.