University at Buffalo
State University of New York

Department of Computer Science and Engineering

February 10, 2016

Ms. Sophie Ruddock
Administrative Officer
Whiteknights House Room 3.21
University of Reading
Reading RG6 6AH, Berkshire, UK.
Re: Promotion for Guy McC. Haworth

Dear Ms. Ruddock and University Officers:

It is my great pleasure to recommend Guy Haworth for promotion to the rank of Associate Professor. I have known him since 1999 in correspondence over the famous Internet Match between Garry Kasparov and the Rest of the World, which was hailed in Michael Nielsen's book *Reinventing Discovery* as a watershed for crowdsourced research. We have been active collaborators since 2007 with several publications in diverse conferences in data mining, computational games, and artificial intelligence. I have the same rank with tenure at the University at Buffalo; my main research field is computational complexity and I also partner one of this field's major weblogs, *Gödel's Lost Letter and P=NP*, with the distinguished Professor Richard J. Lipton of Georgia Tech. To fill out my one-man-band I also serve at-large on the Anti-Cheating Committee of the World Chess Federation (FIDE), on which I've unfortunately just experienced more activity in January than any calendar year previous.

I am commenting on Guy's research and scholarship; I have no input on his teaching except to say that when he presented our joint paper at the Tilburg 2011 congress of the International Computer Games Association (of which he has served as VP), my whole family were visiting us for the US Thanksgiving holiday and were very taken with the urbaneness and clarity and good-humour of his presentation on the live feed. Before I come to the main theme of Skill Assessment, I should begin with Guy's interest in extreme intellectual uses of computers. This is represented first by his review of work on finding the largest provable prime numbers—when that was quite a deal for computers of the 1980s. Then comes his longterm work on chess endgames. The world's largest mathematical proof, dwarfing the 13-gigabyte computation on the Erdős Discrepancy Problem by Boris Konev and Alex Lisitsa (University of Liverpool) which made headlines two years ago, sits on a computer system in Moscow. It tabulates perfect play for all chess positions with up to 7 pieces. It is 140 times larger than the 6-piece tables computed ten years ago which can be compressed within 1 terabyte, which themselves dwarf the 5-piece tables pioneered by Turing Award winner Ken Thompson in the 1980s. To lend perspective, Harvey Friedman, one of the the world's foremost logicians and proof theorists, has been seeking direct programmable access to the Moscow tables since summer—I spent part of yesterday corresponding with him again about news. They include a position in which one side will deliver checkmate in 545 moves but no human master has been able to give the foggiest idea of why. Reasons for academic investigation of these tables include:

- They show mathematical complexity and “digital dynamics” in a natural setting;
- They exhibit many challenges in constructing proofs with computer aids—indeed, Friedman’s interest is in proving checkmates in case-by-case positions that have 8 or more pieces.
- They demonstrate ways in which machine learning goes beyond human mental capability even to understand results—let alone find them.

Well, Guy found a fourth reason by himself, in work of which I was initially unaware when I was confronted by similar contours following the “Toiletgate” cheating-accusation scandal of the 2006 world championship match. I can hark the basic idea back all the way to famous field-tests by Ken Thompson and Donald Michie of Thompson’s mid-1970s tabulation of the 4-piece King-and-Queen versus King-and-Rook endgame. This endgame is generally a win for the side with the Queen, but human grandmasters were often unable to overcome novel defensive tactics found in the tabulation for the side with the Rook (as I was surprised to see when I heard Michie lecture on this at Princeton in 1978). This raised the idea of measuring the skill of human players of all levels by counting how often they make missteps in this endgame and how accurately they follow the shortest winning path when they do succeed. The special point of endgame tables is that they give an absolute measure of the quality of each decision. Thompson and Michie and others never took this beyond qualitative comparisons—showing that stronger players do better in these tests. Guy made it quantitative not merely by correlating results with the standard Elo Rating system in chess but more importantly by defining a mathematical model for the assessment process. Guy’s model progresses the assessment from an initial state of ignorance by Bayesian update. The updates select points in a parameterized space of *probabilistic fallible agents*—that is, agents who deviate from perfect rationality according to a probabilistic distribution over strategies. Although the goal is to find the agent that most closely models the test pattern of a human player, the agent itself is strictly mathematically defined and can be further analyzed and field-tested on that basis.

In my case I needed to model the accuracy and misstep frequency of (presumably) honest human players of various Elo levels over all chess positions, not just endgames. Absent perfect tables I employ a relative criterion judged by computer programs analyzing games long enough to have strength beyond all human players.[1] My model has two psychometric parameters to Guy’s one, uses frequentist methods rather than Bayesian, and incorporates a tuning adjustment to offset a quirk in large data that no one else had observed. Guy, however, had from the start the one vital feature spurned by Ivan Bratko and Matej Guid whose 2006 work limited to chess world champions is usually credited as the first in this line: our models need and use authoritative values for *every* move option, not just the value of

[1] Kasparov lost to the supercomputer Deep Blue in the mid-1990s. Have you heard the one about how an I-Pad or smartphone nowadays is as powerful as supercomputers were in 1994? There’s the problem—I’m writing this just as a master player was caught red-handed using a smartphone last weekend in the top section of the Moscow Open, in a case that hasn’t even reached my committee yet. But it also affords the ability to “fight fire with fire” by using today’s chess programs (running on better hardware) as an authoritative reference in lieu of the tables.

the best move and the value of the played move when it differs. Guid and Bratko's simpler tallying is adequate for relative comparisons of players, but tellingly they have not attempted to create a rigorous correspondence between regressed model parameters and Elo levels *with confidence intervals* as our models do. That's what is needed to speak of Skill Assessment in an absolute numerical sense. Then we can reap the major benefits:

- Your assessment is based directly on the judged quality of your individual decisions, whereas Elo ratings are based only on the final result of the game (win/lose/draw) and hence subject to fortune of the opponent's good or poor play.
- Whereas master and amateur players both typically play only 50–100 games in competitions per year, a terribly small sample in terms of game results, their individual moves in those games (even after excluding the first eight or so opening moves) will give a healthy sample of 1,500–3,000 relevant moves.
- For my anti-cheating vocation, the confidence intervals support null-hypothesis testing and other statistical judgments—where unlike the famous cot-death cases I am able to field-test the statistical judgments.

Although Guy's initial work in 2002–2003 and 2005–2007 used only the win/draw/loss values and progress toward checkmate numbers from the endgame tables, his model adapted readily to the more general game setting with values supplied by the strongest chess-playing programs. It must finally be emphasized that these values are the *only* features of our models that pertain to chess in any way. Our models can thus carry over to any game or any setting in which decision options can be given authoritative (or objective "hindsight") values, including the scoring and design of multiple-choice standardized tests.

Now to review our joint papers: Guy and Giuseppe DiFatta were first out of the gate programming the one-parameter Bayesian model. I joined them as third author in a successful paper to the 2009 IEEE Computational Intelligence in Data Mining symposium and then in a more chess-specific paper for the 2009 ICGA congress (the Springer LNCS proceedings as final publication appeared in 2010). Meanwhile I charted a longer flight path for my more-complicated model. This plus a more-selective core training set and deeper analysis with a better program (Rybka in place of one called Toga II) supported a positive demonstration that—contrary to widespread belief of Elo ratings have "inflated" relative to absolute skill—the rating system has been remarkably stable since its international inception in 1971. This paper was accepted to AAAI 2011, considered along with IJCAI the world's premier conference on artificial intelligence. This was followed up by our 2011 ICGA (Tilburg) paper showing other measures to corroborate the no-inflation finding, on which we were joined by a strong grandmaster player from Poland who has advised FIDE on Elo-rating matters. My graduate student Tamal Biswas joined my work in 2011 and took it more toward machine learning, psychometrics, and decision theory (you can find this in a GLL blog article titled "Depth of Satisficing") while I was pressed into anti-cheating service following the breaking of the Borislav Ivanov case at New Year's 2013. But Guy has been a close advisor all this time—he took several days to act as "Dutch Uncle" to mellow my open letter about the

Ivanov case—and we three got back together for a critical review of anti-cheating methodology which he presented at ICGA 2015. This paper includes our response to a notable article by David J. Barnes (of Java textbook fame) and Julio Hernandez-Castro in the mainstream journal *Computers and Security* a year ago.

All of this is besides Guy's separate academic work on plagiarism detection (whose details I am not qualified to go into) and his other papers on chess endgames. I had desired to join the "Position Criticality" paper with work on some novel mathematical issues it raises, and Lipton and I have taken a different tack of auto-fixing errors in critical positions (blog post: "Playing Chess With the Devil"). This goes to say that Guy has raised in the chess context some questions with longer legs. All of this work is marked by decidedly careful scholarship mindful of the full historical context—you can see this in our jointly-created presentation slides at `http://www.cse.buffalo.edu/∼regan/Talks/ReganMaciejaHaworth2011.pdf`. All he would need to bring this up to thoroughbred is the penchant to sleep in the stable with the horses—which also means sometimes not sleeping—or have assistants who can. The last is what either of us would need to try to make this work drive a chariot of assessing multiple-choice examinations for intrinsic difficulty and/or detecting plagiarism in item-response settings. Incidentally, we finally met in person for the first time last June at Reading and November in Oxford, where I presented much of the above to the Computer Science Department.

In all, Guy has been extraordinarily active in publications and reviews, let alone that he returned from industry to an academic appointment that has been mainly for teaching. The skill-assessment idea via Bayesian update is a quintessential original core of a doctoral thesis, and there ought to be some way that a 50-page collecting of his early "reference-fallible" papers and the 2009 Bayesian model plus our other joint work could lead to conferral of the degree. I've underscored in this letter how his work connects to computing and mathematics on the whole. I will be happy to answer any more-particular questions; meanwhile I trust that this suffices to attest the requirements of the promotion.

Yours sincerely

Dr. Kenneth W. Regan