University at Buffalo
State University of New York

Department of Computer Science and Engineering

May 19, 2016

Lecture Recruiting Committee (Data Sciences)
Goergen Institute for Data Science
734 Computer Studies Building, Box 270226
University of Rochester
Rochester, NY 14627
Re: applicant Tamal Tanu Biswas

Dear Lecturer Recruiting Committee:

I am delighted to provide a reference for Mr. (soon to be Dr.) Tamal Tanu Biswas. For five years he has been an indispensable partner on my research on human decision making at chess and driving force in connecting to wider areas of computing and human behavior. He has generated creative ideas on which to build a research program, brought them to fruition in published papers (some solo), and proved himself able to work energetically with multiple people and be an excellent colleague and teacher. His research described below incorporates AI, machine learning, decision theory, and computational science with big data. Among CS graduates he has a winning combination of ability to code software systems that work, conduct his own experiments with millions of data points, and formulate innovative ideas to test.

I realize yours is not a research TT position. I've decided to use much the same letter as for TT because it's all relevant for Tamal's diversity of ability. Besides saying I can vouch for his teaching statement, I can add that he taught in summers with *no* oversight from me—I was too busy—beyond some initial advice since I had taught the same course (CSE250, Data Structures in C++) in the spring regular term. I say more about his teaching and personal qualities at the end.

For myself, most of my academic career has been in computational complexity in the theory-and-algorithms area. I also now partner the widely-read weblog *Gödel's Lost Letter and P=NP* (GLL) started by Richard J. Lipton, with whom I've done research and co-authored a textbook on quantum computation published a year ago by MIT Press. I was a junior chess champion and hold the title of International Master from the World Chess Federation (FIDE), but always turned down suggestions to join academic work on computer chess. That is until scandalous accusations broke at the Sept.–Oct. 2006 world championship match that the Russian player, Vladimir Kramnik, was cheating by consulting the program Fritz 9 backstage during the games. Fritz 9 would shortly demonstrate its superiority over all human players by beating Kramnik in a match with no defeats while running on ordinary home-PC hardware, not a supercomputer as with Deep Blue over Garry Kasparov, and today's programs are so much better they can trounce us on smartphones. The only evidence given was statistical, and I answered a request by the program's maker for help in how one can possibly evaluate such accusations.

I first built a qualitative model around the principle that if you have few reasonable options then your likelihood of agreeing with computer programs goes up, demonstrated how the challenger had been unusually forcing Kramnik's hand to explain the parts of the accusation that were reproducible, and used the model unofficially in a few cases in 2007. Then I began expanding the model quantitatively and gathering data from the voluminous public record of games by players at various skill levels measured by the venerable *Elo rating system*. By summer 2010 when I met Tamal I

had over 400,000 analyzed positions in my main training set covering eight Elo century levels and the years 1976–79, 1991–94, and 2006-09, plus 600,000 more positions from top-level events of all years. The games at lower levels needed greater manual cleansing of gamescore errors, 56 errors from 6,000+ games in all, several of which had made it seem the players were leaving their Queens in take on every move had the recording errors not been fixed. I had 10,000 lines of C++ code for statistical analysis, nearly 2,000 lines of Perl scripts for collating the raw data from programs, and a full working *predictive analytic* model that generates not only a wide range of statistics but also confidence intervals for hypothesis testing.

Statistical cheating detection is a real-world problem, and in chess it needs "fighting fire with fire"—using chess programs to generate values and infer probabilities by which to detect fraudulent use of (other) programs. The use of the programs' final values for *all* available moves in each position distinguished mine from prior models by Ivan Bratko and Matej Guid and others, all except a Bayesian model by Guy Haworth of the University of Reading, UK, with whom I've joined forces since 2009. Beyond that I kept it minimal. The model still today uses *no* feature of chess except the move values. The minimalist design averts potential player-specific bias in cheating cases and allows transferring the model to other games and decision processes. Haworth and I swapped authorship on our respective models and the paper with mine was accepted to AAAI 2011.

Tamal's Early Project Work

In fall 2010 I showed Tamal the ropes as he finished his Master's in our sister EE Department having secured admission to my department's PhD program, which he deferred to August 2011 to return to Bangladesh for community and spiritual service. Then in January 2011 came the first major cheating case involving a top-100 player, Sebastian Feller of France, and I was retained by the French Chess Federation prosecuting it. Working from Bangladesh, Tamal programmed and analyzed multiple suites of 10,000 resampling trials from the training data as empirical tests that the *z-scores* produced by my model conform to normal distribution. As the case generated multiple lawsuits I was glad he was less exposed there than I was here, and his work stood up when we ran an upgraded validation suite in summer 2014. Thus he was in the thick of things by fall 2011.

My infrastructure goals in 2012 were upgrading to a richer data format, harmonizing values given by multiple chess programs, and enhancing the model by exploiting data recorded at *lower* depths of search before the programs' gave their final values. This all would multiply the volume of data by upwards of 100, straining running times already being measured in hours. Before we even got to that stage, several scientific and software obstacles emerged:

1. Adding a parameter d for "depth of thinking" to the model—via equations involving the lower-depth values that I'd sketched in the future-work conclusion to the AAAI 2011 paper—turned out to convert a previously smooth fitting landscape into one pitted by bad local minima and a perverse global drift when regressing for d.
2. The multiple programs—except one called Houdini—obey a logistic-curve law discovered and advocated by Amir Ban (who is better known as a co-creator of the USB flash drive than the chess program Deep Junior) which allows putting their disparate evaluation functions on a common scale. But their "second moment" behaviors differ so greatly (buzzword: heteroskedasticity) that I've recently scrapped the idea of combining their outputs.

3. No public APIs or scripts existed for the finer GUI-free control we needed of multiple chess programs under the common protocol called UCI. (The only one since then, `python-chess` by Niklas Fiekas last year, does not preserve lower-depth data.) Designing and settling the new file format also took time as we nervily rejected both XML and JSON in favor of consistency with chess standards. Ours is described in a GLL post titled, "A Computer Chess Analysis Interchange Format."

Tamal did extensive programming for all three goals. He added 2,500 lines of C++ to handle the new data format and program exact numerical integration for my apparently-novel "Percentile Fitting" method in the AAAI 2011 paper. The latter helped me see with larger data that simpler methods were just as good in all respects and better in some. Automating the task of flagging gamescore errors and generating legal moves for further analysis turned out to be prohibitively slow with the "naive" routines I'd found in the public Perl domain, but faster code took us fully into the task of writing the *full third* of a chess program that handles board representation and efficient move generation where none existed in Perl. In `python-chess` the module for this is almost 4,000 lines. Tamal eventually hit on the idea of leveraging the needed code in the public-source Stockfish chess program by writing a Perl-C++ SWIG interface. This has served us well, but in adapting our code for portable use by FIDE, last year I wound up completing and trebling the 1,300 lines of pure Perl that I'd started anyway. In the meantime we manually cleansed a new data set of 726,120 moves from just over 10,000 games representing the years 2010–12, for each of four programs. Finally, in 2014, Tamal finished control scripts that enable us to run in batch mode on UB's supercluster, so we've started running the entire 7+ million game corpus of recorded chess history; we can't hope to data-cleanse them all but we now know enough about the incidence of error to make reasonable automatic corrections.

Thus in 2011-12 Tamal shouldered responsibility right away and acquired proficiency in Perl and C++, which he's put to further use teaching our department's required object-oriented C++ course in the past three summer sessions. He handled large-data experiments and wrote scripts in Perl and MATLAB and learned the latter's tools for generating figures. Then at New Year's 2013 the (in)famous case of Borislav Ivanov broke and presaged multiple cases in 2013 that commanded my involvement and led the World Chess Federation (FIDE) to take unprecedented action. To finish my story: I served formally from 6/13 to 12/14 on a joint committee of FIDE and the Association of Chess Professionals to draft new regulations and guidelines to combat cheating; I remain a chief consultant to its current FIDE incarnation. Calendar year 2014 also occupied me throughout on the textbook with Lipton, while 2015 brought a horrible spate of cheating cases and false accusations—and 2016 has opened even worse with *four* new cases (two official) since New Year's Eve. But now to tell the core of Tamal's story.

Tamal's Research

In Spring 2013 I offered an extra entry-level graduate seminar on the chess research, which Tamal formally took but really acted as assistant to several groups of newer students on their required projects and presentations. My attitude with that seminar was that "you guys should teach *me* how methods learned in your courses on machine learning and AI and pattern recognition and computational science can be applied here." I started with a month of demonstrations and lectures on the basics and open problems to investigate and experiment on. This included the correspondence between a chess position with k reasonable moves and a multiple-choice question with k

options, expanded further along lines of a post on the "Angry Statistician" blog amusingly titled "Baseball, Chess, Psychology, and Psychometrics: Everyone Uses the Same D— Rating System" which appeared during spring break.

Tamal and I packaged these correspondences and the above procedural refinements into a submission to the IEEE Computational Intelligence in Games conference. Then in May after the term ended, Tamal presented to me what he had done with the correspondence to *Rasch modeling* and *item response theory* (IRT), theories used among other things to evaluate and set baselines for standardized tests and corporate personnel-evaluation tools. His drift is how to "judge the judge" for these baselines. I was really impressed. Our CIG 2013 paper was already accepted and we did not find either propriety or page-room for adding his deeper material, so I remained first author of it. Unfortunately a Canadian embassy strike prevented his obtaining a visa to give the talk even though CIG 2013 was just across the border in Niagara Falls, Ont. Eventually this material grew into his solo regular paper at the January 2015 International Conference on Agents and Artificial Intelligence (ICAART 2015), which he presented in Lisbon along with our joint submission described in detail below.

Then we set about tackling the 100x data blowup problem. We hit on the idea of precomputing (or memoizing on the fly) values of an expensive high-dimensional function f at the heart of our regression loops at selected nearby gridpoints, then interpolating them. Using Taylor approximation has the downside of needing to (pre-)compute at least the first partial derivatives, but they are too many to store. Moreover the distribution D of arguments $\vec{x}$ to f, being derived from columns of values given by the chess programs (at each depth of search), is far from uniform. D obeys properties, however, that seem to allow finessing the needed partials by a balancing strategy that also involves a one-pass heuristic for cases of the knapsack problem.

Tamal designed a family of data structures—trees with graduated branching and an efficient code for locating branches—that both emulate D to build the nonuniform grid and facilitate the knapsack balancing process. The implementation and experiments were also all his. The results were superb on randomly synthesized large-scale data but a wash on our actual "pre-blowup" small-scale data—while the landscape issues mentioned above then superseded our plans to tune it on the large scale. Our paper was accepted to the July 2014 AAIM (Algorithmic Aspects of Information and Management) conference in Vancouver. For schedule not visa reasons I gave the talk there while Tamal presented another paper[1] at the Multidisciplinary Preferences workshop associated to AAAI 2014 in Quebec City three weeks later. The Vancouver audience included Mario Szegedy, famous as a progenitor of *streaming* algorithms for handling large data, and he told me afterwards he really liked the ideas. Our paper was invited to the conference special issue and has appeared in the

[1]This paper, which included a high-school student now at the University of Chicago, used scripts written by Tamal and the other student to analyze the new 4x726,120-move human-play data sets, and those plus the C++ code on other data involving machines playing alone and human+computer "centaur" teams, to explore differences between human and computer decision phenomena and preference "styles." The results include demonstrating that humans but not computers have lower expectation when it is their turn to move, that humans play in a more forcing style than computers, that a psychological explanation s favored over a rational one for an error-scaling phenomenon distinctive to human play, and that the observed superior results of centaurs over computers playing alone in a high-level series of "Freestyle" tournaments played in the years 2005 to 2008 were really due to finding intrinsically better moves—as judged by the superior chess programs of 2014 against which the effect is no longer found. The chess "centaur" motif has recently been adopted by the US Department of Defense at highest levels for the "Third Offset" strategy of human+computer command and control. I was properly first author, but Tamal wrote all the slides and presented well.

major journal *Theoretical Computer Science*.

While I was diverted by finishing the quantum book and time-consuming vicissitudes with FIDE that climaxed on a bizarre and tragic final day of the 2014 Chess Olympiad, Tamal revisited our depth-of-thinking idea. Rather than make a new model parameter d for depth in the equations, he conceived that the by-depth columns of data should first be synthesized into one or a few scalar measures, each denoting a concrete *effect*. Then the effects can be meaningful adjustment factors on the equations. Tamal adapted his measures from existing literature in Rasch and IRT and machine learning and justified adopting standard names for them, in particular *discrimination* and *difficulty*. He also settled on a formulation of a move's "swing" in value across depths. The sensitivity of his measures was a shock to me. For instance, on positions in top-level games that register between 4 and 5 on his 0–5 scale of "swing" the observed move-match statistic plummets from the overall 58% to 30%, contrasted with 70% in the 0-to-1 interval. This difference occurs even though the top-depth columns alone are substantially the same for the two sets of moves. Hence this phenomenon was being missed by the "pre-blowup" setup—although on sets of 1,000 or even just 10 games the kinds of moves usually mix enough to even things out. This became our joint paper at ICAART 2015 in Lisbon. With travel funds and assurance-of-visa in hand he ventured to write up his "judging" material along with a smaller submission to the "doctoral consortium" day. Neither had writing-review from me on the scale of "teaching him how to write" from the two previous years; a few phrases and figures could have been better polished but they were fine according to the referees who accepted them.

Then in spring 2015 Tamal created a *concept*, fusing depth with Herbert Simon's durable concept of *satisficing*. Satisficing means being satisfied to work toward a goal that suffices rather than aim for an optimal return. We might always think we make an optimal decision when we stop thinking about it, but the kicker is how much more thought and time and effort we were (not) willing to invest in the quest for a better one. The innovation is to find common units for diverse notions of "time and effort" in the measure of *depth*. Once these units are in hand—in any model not just chess where depth-of-search provides a ready formulation—one can quantify for each case k the amount of (additional) depth d_k that would have been required to reject the chosen decision in favor of one that proves to be better.

In the chess model, the depth d_k is one more than the last depth at which the played move was optimal (if any). This is most often the same as the depth at which the computer's eventual best move overtakes the played move's value. Tamal uses a third criterion that applies to more cases and is numerically smoother to work with, but in all common cases the values are close and the interpretation is the same: it is the depth at which the player's error is exposed by the superior analysis of one or more computer programs acting as judges. The average d_P of d_k over all positions k where the player P's move "swings down" becomes that player's average *depth of satisficing*.

The neat result is that d_P has a strong linear correspondence to the skill rating of the player P. The regression showing this uses only the swing-down moves. That is, we are able to infer a player's skill solely by looking at cases in which the player screwed up. That deeper players make deeper mistakes might not be surprising from a Bayesian standpoint, but the strength of the linear fit and the fact that d_P starts at 10 for the world's elite (no higher) and extends all the way down below 3 both indeed surprised me. The paper is mostly his and he presented it last month at the 17th IEEE International Conference on Machine Learning and Applications (ICMLA, not to be confused with the more prestigious ICML) in Miami. I described it more technically on the Gödel's Lost Letter

blog:

https://rjlipton.wordpress.com/2015/10/06/depth-of-satisficing/

We have no shortage of things we can do. Highest on the purely-chess side is working out the (chess program-dependent) exact adjustments for the prediction equations from Tamal's demonstrated swing effect and hopefully sharpening the efficacy and conformance of the cheating tests. For potential wider application I should emphasize that this extension is still using *only* the numerical values given by the analyzing programs, no other relation to chess at all. Last year Tamal and I also wrote a paper with Haworth surveying the general issue of using chess-dependent factors and profiling individual players' prior tendencies in cheating tests. Examples of the former are whether good retreating moves are harder than usual for human players to find and whether players fixate on *plans* to the detriment of later decisions that should be considered afresh. Both are commonly asserted but neither has been subjected to thorough quantitative testing. How large an effect do they have on prediction for human players? I could go on, but it is more important to say what research this has enabled Tamal to do apart from chess.

Tamal's Future Research and Qualities

One major area is taking the theoretical IRT correspondence into the field. A first-moment question is, how can we vet standards for what should be an A, B, C... in multiple-choice tests for large-scale online courses? Higher-moment questions include: How can we tell how sharply a test question discriminates between these grade levels? How much depth of reasoning is required to find the best answer? The first advantage of chess that a correspondence can leverage is that the quality standard given by the Elo Rating system has been remarkably stable internationally since its inception in the 1970s. My earlier work showed the absence of FIDE rating "inflation"—against conventional wisdom but in line with what analogous population fitness models predict. Tamal's work has extended this advantage to higher moments where having large data is singularly important to establishing relationships. A higher-level correspondence will provide grounding for test-taking analytics and identify population phenomena that transfer and hence may be targeted in test design.

Within decision theory our work has been criticized on grounds that the population of chess players is specialized and chess is not targeted to specific social-choice phenomena the way specially designed games in the literature are. This leads into classic "big-data" arguments over mining for discoveries versus testing theories. We've tried to bridge this by saying that the lower specificity but higher data from chess confer commensurably valid support for some theoretical positions over others. The chess work also focuses on what appears to be a distinctive problem of inferring probabilities of actions from utility values that are provided by an essentially perfect judge but perceived only approximately by the actors. This connects for instance to financial fields where success at sniffing out the future is paramount and not judged by aptitude testing for knowledge and quantitative ability.

Tamal has a higher ambition which definitely requires lifting off the launching pad of chess: to analyze the weighing of multiple criteria in decisions. Some of the multiple-criteria literature focuses on simple binary cases where one criterion is minimizing time/cost, which can be modeled by studying time budgeting and pressure in chess, but the most interesting problems stand apart.

Tamal is branching his work into areas where important datasets will be more readily available. The experience I've described had given both great training and perspective on working with data, including the creative frustration and scientific maturity that come from unlocking truths in cases where the data doesn't initially "behave."

Fraud detection is another area jumping off from the chess cheating work. I would be pursuing applications to cheating in massive online games except that my own game has unfortunately been filling all time. Tamal has witnessed some major statistical pitfalls as well as tools. The flip side of fraud detection is information assurance and the design of robust systems. This is where his work with Professor Upadhyaya on security and safety of networked systems is a valuable counterpart. Their paper was just now accepted to the flagship conference of the IEEE Vehicular Technology Society, VTC2016-Spring in Nanjing, China.

Tamal's greatest immediate push, however, will be melding his take on the concept of *depth* into machine learning. Why is greater depth needed in neural networks? In abstract complexity terms, this is pushing the constant depth threshold circuit class (called uniform TC^0) toward higher levels of the polylog circuit depth/parallel time hierarchy (called NC) or even polynomial time (P). Tamal is not out to do complexity theory like myself, but does have the grounding in theory and algorithms to appreciate the impact of depth architecture on problem solving. We have a recent prod to do this in chess from the high-level attention given to the recent development of a deep-learning chess program called *Giraffe* that acquired my own level of chess ability (Elo 2400) by 72 hours of learning. Some of this attention is coming through the same channels as interest in the "centaur" initiative by which it proceeds into human-computer interaction. Using our model and data to track where and why this new kind of chess program works, and thereby identify success indicators for other applications, may round out his dissertation, which is already burgeoning with his work to date.

In sum, he is building a distinctive research program, already branching apart from his advisor, and will be a valuable member of a research community in which diversity of knowledge and application is increasingly important (including for external funding). He is one of the hardest-working people I've known and has taken successful initiative in research while I've been diverted. He has not TA-ed any of my classes—he was pulled away from doing CSE250 under me in spring 2013 by need in another undergraduate course, and has continued to serve courses with high undergraduate interaction in labs. He is an excellent and organized lecturer; he did learn that slowing his delivery modulates his subcontinental accent. Most of all he is a pleasure to work with, including some stubbornness that commands respect as some code-architecture decisions I initially demurred from turned out to be best. He also has a highly mature sense of restraint, and I can simultaneously say that some of this both comes from his discipleship and is indicated by his decision *not* to say anything about this on his application materials (for "diversity" or whatever). This aspect has played into a deep personal friendship with me and also ready approachability with others. All of these are strong indicators of success in a collegial environment, which is why I am promoting tenure-track applications without the intervening stage of a postdoc. I can give him my highest recommendation and will be happy to answer any further questions you may have.

Yours sincerely,

Dr. Kenneth W. Regan