Detection Of Non-Human Agents

Richard J. Lipton and Kenneth W. Regan

In modern complex distributed computing systems one of the recurring issues is: Are we interacting with a human? a computer? or a combination of a human aided by a computer? Determining which is the case is often of paramount importance in securing a system or understanding what is happening. All sources of improvements in the state of this art carry vital potential.

In **active** interaction, one party, Alice, is able to challenge the other party, Bob, with questions. This is essentially a type of Turing test, and is relatively well understood. Accurate and powerful protocols have been developed, for instance CAPTCHA technology based on visual images.

We will study the **inactive** case, where Alice is restricted to watch Bob's behavior and *detect passively* the event that Bob is (not) human. This problem is fundamental and uniquely challenging. A solution could advance security, social media data mining, and other basic technologies.

The plan of research is novel in branching via a well-defined problem: Is a player of a chess game cheating—that is, using a computer to make his moves? Today even the world's best players are rated hundreds of points below chess engines that run on laptops. This state has led to many real-life cheating cases, where players in high-level tournaments have been accused and sometimes found to have cheated. The second PI has developed a statistical model of human versus computer move choice in papers since 2011, and has been involved (as "Alice") in several such cases. The model takes deep computer analysis of all move options and player skill parameters as inputs, and generates probabilities for each move and predictions for aggregate performance over sets of moves.

This adds a third detection category, since Bob is a *human using a computer tool*. Special chess competitions called "Freestyle" allow human-computer collaboration. This may seem hard to separate from "human" and "computer," but a new joint paper by the second PI, to be presented at the M-PREF workshop associated to AAAI 2014, does exactly that in the case of chess. This supplements move-choice statistics by measuring criteria by which humans tend to force games into earlier crises, while computers acting alone direct games into positions where they have more reasonable options and can play wait-and-see.

Applications will also come from the correspondence, established by the second PI's joint paper at IEEE CIG 2013, between chess positions and multiple-choice test questions, and between the mathematics of rating chess players and theories of **psychometrics** used for personnel assessment. The important point is not cheating and chess *per se*, but rather the ability to leverage huge amounts of data—millions of moves from real competitions not simulations—in which mathematical tools such as the Elo rating system and logistic analysis are fully established. This work has already led to discovering "big-data" style regularities of human behavior, including that human error scales up with the magnitude of either side's advantage in a position, and mass sensitivity to small differences in *post-hoc* move value.

Thus we claim that using a real concrete problem—chess—to help develop a theory of passive detection of non-humans is important, with promising transfer potential to applications beyond chess. The new proposed research will add Bob's decision **depth** and **time** as primary factors, ones apparently absent in Rasch or item-response psychometric theories, and thereby deepen all of the agent-discrimination, distributional performance-assessment, and test-difficulty rating components. Anticipated project duration is 36 months at 10% term, 50% summer effort.