The Predictive Mind and Chess-Playing. A Reply to Shand (2014)

Matteo Colombo and Jan Sprenger

Abstract In a recent *Analysis* piece, John Shand (2014) argues that the Predictive Theory of Mind provides a unique explanation for why one cannot play chess against oneself. Based on this purported explanatory power, Shand concludes that we have an extra reason to believe that PTM is correct. In this reply, we firstly rectify the claim that one cannot play chess against oneself; then we move on to argue that *even if* this were the case, Shand's argument does not give extra weight to the Predictive Theory of Mind.

Keywords: Predictive theory of mind; Chess; Explanatory power; Inference to the best explanation

In a recent *Analysis* piece, John Shand (2014) argues that the Predictive Theory of Mind (PTM, henceforth) "neatly explains why one cannot play chess against oneself" (p. 5). Based on this purported explanatory power, Shand draws an inference to the best explanation to conclude that we have an extra reason to believe that PTM is correct.

In what follows, we firstly rectify the claim that one cannot play chess against oneself; then we move on to argue that *even if* this were the case, Shand's inference does not give extra weight to PTM. We note that our conclusions can be generalized to other cognitive phenomena.

According to Shand, it is impossible to play chess against oneself because "uncertainty of what one's opponent will do, and the subsequent positions, are essential to playing the game" (p. 5), and in self-play one cannot be uncertain about one's own next moves: "while thinking of one's own move it is simply impossible not to know what one is very likely going to do in reply to it" (p. 4). Thus, chess is essentially represented as a prediction game in Shand's argument: "Chess is about thought battling uncertainty and the millions of possible permutations and variations in, and extending from, the positions chess involves" (Ibid.).

First, we observe that many chess players, including grandmasters as well as one of the authors of this reply, do sometimes play against themselves. The famous grandmaster and vice world Champion David Bronstein (1924-2006) even played a beautiful game against himself while sleeping, becoming famous as “Bronstein’s Dream Game” (see Appendix). This game also has some significance from the point of view of chess opening theory. But Shand (2014) dismisses these cases as irrelevant by noting that, there should be more to a game of chess than to moving the pieces according to the rules of chess (p. 3).

This claim brings us to the psychology of chess. If the argument for why you cannot play chess against yourself is that no predictions about what “your opponent” is going to do affect self-play (since there is no uncertainty about one’s own moves), and predictions about the other’s moves are necessary for playing chess, then psychological evidence offers thin grounds to conclude that self-play is impossible.

There are several kinds of cognitive processes that can be involved in playing a game of chess—including problem-solving, acquiring and retaining in memory a large number of patterns of chess configurations, chunks and templates, pattern-recognition, quick pattern-retrieval, evaluation, counterfactual thinking, and, of course, look-ahead search and anticipation of the opponent’s most likely moves (cf. Chase and Simon 1973; Holding 1992; Gobet and Charness 2006). According to some prominent theories (Chase and Simon, 1973; Gobet and Simon 1996; Saariluoma 1995), chess expertise taps most crucially onto quick and reliable recognition of valuable board configurations; processes such as search, calculation of variations, and brute-force prediction of the opponent’s most likely next moves play a far less important role. Interestingly, this fact neatly distinguishes human from computer chess players. As recalled by the Russian chess grandmaster Garry Kasparov, “correctly evaluating a small handful of moves is far more important in human chess, and human decision-making in general, than the systematically deeper and deeper search for better moves—the number of moves ‘seen ahead’—that computers rely on” (2010). Such evaluations include the understanding of pawn structures, the recognition of “harmonious” piece configurations and the intuitive decision-making that is typical of blitz games (five minutes per player and game). If this is correct, then there is little justification for the belief that predictive skills are necessarily engaged in all instances of chess playing, even at the expert level. But if this belief is unjustified, then Shand's argument for why one cannot play chess against oneself rests on thin grounds.

Games against oneself are not only psychological possible, but also non-trivial in various ways. Playing White, I may face a difficult strategic choice (e.g., between two different pawn structures), while playing Black, my moves happen to be more or less forced. The predictive dimension of games against oneself need not be trivial (or non-existent) either. Playing White, I spot a sequence of forced moves, which appears favourable to me. However, my judgment may have been mistaken: the judgment of the final position was too optimistic, or I might have overlooked a forceful reply for Black somewhere in the combination. Only when the forced moves are executed, that possibility transpires to both White and Black because they are now in a better position to visualize the mutual opportunities and the final position. Such phenomena occur regularly in self-play *and* in tournament games between two players (Krogius 1983).

These considerations put pressure on Shand's claim that one cannot play chess against oneself. Of course, *some* games between two players could not occur when one player is battling oneself. Neither do we claim that there is no meaningful difference between real and imagined games. But this does not imply that playing chess against oneself is absurd or psychologically impossible. Our conclusions can be easily transferred to other games with complete information, such as Go, checkers or backgammon.[1]

Now, in the second part of this reply, we show that *even if* Shand were correct about the crucial skills in chess-playing, the purported inability of playing chess against oneself could be explained in different ways.

[1] In computer programming, reinforcement learning from self-play is one of the most interesting approaches for studying and training computer programs that can play complex board games—including Go, backgammon, checkers and chess. In this approach, the program plays many games against itself, keeps track of sequences of board positions starting at the opening position, and uses a reward signal obtained at the end of each game to improve the quality of its move decisions in subsequent games. Although self-play is a time-consuming learning approach, where it may be relatively hard to detect which moves are bad, it can lead to the same level of performance of expert human players (cf. Samuel 1959; Tesauro 1994; Thrun 1995).

Generally, explanations of why a certain kind of system cannot produce a certain cognitive phenomenon should provide us with information about what cognitive resources the phenomenon in question recruits, and why the system lacks, or is unable to deploy, those resources. Specifically, explanations of why human agents cannot play chess against themselves should provide us with pieces of information about what cognitive resources chess self-play recruits and why the human cognitive system lacks, or is unable to deploy, those resources.

In order to explain the inability to play chess with oneself, Shand (2014) firstly sketches a very general theory of brains, cognition, and life indeed, viz. what he calls "the Predictive Theory of Mind." He then describes chess-playing in terms of general concepts of this theory. Finally, he tries to show why it is absurd to apply the same concepts to characterise self-playing in chess. Thus, our inability to play chess against ourselves would rest explained, and we would have an extra reason to believe that the PTM is correct, or so Shand claims.

What is the PTM? And to which claims is it committed? The PTM is most closely associated with recent work by Karl Friston (2010), Jacob Hohwy (2013) and Andy Clark (2013). The basic commitments of the PTM are twofold: first, brains are kinds of prediction machines; second, brains produce cognitive phenomena, including perception and action, by constantly attempting to minimize prediction errors.[2] Prediction errors quantify the discrepancy between predictions about the value of a certain variable and the observed value of the variable (Niv and Schoenbaum 2008). In the PTM, prediction errors quantify the mismatch between expected and actual sensory input. Specifically, according to the PTM, the brain would encode models of the causal structure of the world. Based on these models, it attempts to predict its sensory inputs. If the brains' predictions about sensory signals are not met, then a prediction error is generated, which will tune the brains' models of the causal structure of the world so as to reduce the discrepancy between what was

[2] It is not obvious what the explanatory scope of the PTM is. Shand (2014) suggests that any cognitive phenomena can be accounted by the PTM when he writes that "the brain in encountering chess is only doing what it does when dealing with *any* experiences and determining how to act" (p. 4, emphasis added).

expected and what actually obtained. Friston, Hohwy and Clark agree on these two basic tenets, as well as on how these tenets can be most properly fleshed out in technical terms borrowed from information theory and statistics, viz. in terms of *free-energy*, *Bayesian inference*, and *hierarchical predictive coding*. Despite this agreement—it should be noted—Friston's Hohwy's and Clark's formulations of the PTM are not equivalent, and the consequences they draw from it are different.

Shand conceives of chess-playing as a prediction game where, if you played against yourself, your brain's model of the game could not possibly make mistaken predictions about the next configuration. No uncertainty would need be minimized, as no uncertainty would be involved in predicting your "opponent's" moves. No surprises would be involved. No prediction errors; no game of chess could be played.

It should be apparent that there is no claim specific to chess in this explanation, besides the assumption that genuinely playing chess requires some uncertainty about the next move of one's opponent. Surely, the information-theoretic and statistical concepts embedded in the PTM provide us with one possible quantitative, encompassing, framework whereby we can describe and study behaviour and cognition. However, it is far from clear how such description would explain or provide understanding on particular kinds of adaptive processes, cognitive phenomena and behaviours. Within the PTM, 'prediction' 'expectation' and 'surprise' are information-theoretic notions that should be neatly distinguished from the corresponding psychological concepts that we ordinarily employ to pick out personal-level mental states and explain particular behaviours. Furthermore, describing *any* cognitive phenomena in terms of the information-theoretic and statistical notions embedded in the PTM affords a kind of unification that is not obviously explanatory (Colombo and Hartmann ms).

Although Shand claims that "it is hard to see [the inability to play chess] explained in any other way" (p. 5), there are alternative frameworks where the target-explanandum can be described, studied and *possibly* explained. Alternative frameworks that readily come to mind include game theory and reinforcement learning (Von Neumann and Morgenstern 1944, Sec. 15.7; Sutton and

Barto 1998; see also Shannon 1950). Here is one *possible* sketchy explanation of why humans are allegedly unable to play chess against themselves, cashed out in terms of game theory and reinforcement learning.

One basic model of chess is that of a Tree Search problem, where each state is a particular board configuration and the available actions correspond to the legal moves available to a player at a certain time. Looking forward to the end of the game and apply backward induction to find a solution is intractable. A player needs an *evaluation function* which takes as input the current state of the game and outputs an estimate of the goodness of that state. As the ordinary goal of a game of chess is to win the game, the goodness of a state is defined in terms of the expectation (or likelihood) to win the game making some move from that position. Given an evaluation, the player needs to rely on action-selection policy or some heuristic to decide which move to take.

Now, one possible explanation for why, in some sense, you cannot play chess against yourself is that in self-play your evaluation function will define a goal different from the ordinary one of winning the game. If you and I play chess against one another, our evaluation functions will define our goal, which, ordinarily, is to win the game. During the game, we shall make our moves so as to maximize our likelihood to win the game, thereby reaching our goal. If you play against yourself, your evaluation function could differ, leading to different implicit goals. You may just aim at exploring certain branches of the Tree, for example those that lead to bizarre, sharp or exciting positions. But if your evaluation function defines a goal that is not to win the game, then, in some sense, you are not playing a game of chess at all. That's why, in some sense, you cannot play chess against yourself.

The point of this sketchy explanation is to show that there are alternative frameworks, whereby the target phenomenon (i.e. the inability to play chess against oneself) can be described, studied, and possibly explained. Shand's inference to the best (or only) explanation, which he takes to be the PTM, is undermined by the introduction of a new competitor, even in the absence of any new data. If there are frameworks alternative to the PTM whereby we can describe, characterize and possibly explain the target explanandum, then, whatever the vices and virtues of the PTM, Shand (2014) has not given us an extra reason to believe that the PTM is correct.

References

Chase, W. G., & Simon, H. A. (1973a). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science, *Behavioral and Brain Sciences* 36(3), 181-204.

Colombo, M., & Hartmann, S. (unpublished manuscript). Bayesian Cognitive Science, Unification, and Explanation.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Review Neuroscience,* 11, 127-138.

Gobet, F., & Charness, N. (2006). Chess and games. Cambridge handbook on expertise and expert performance (pp. 523-538). Cambridge, MA: Cambridge University Press.

Gobet, F. & Simon, H. A. (1996). The roles of recognition processes and lookahead search in time-constrained expert problem solving: Evidence from grandmaster level chess. *Psychological Science*, *7,* 52-55.

Hohwy, J. (2013)*, The predictive mind*, Oxford: Oxford University Press.

Holding, D.H. (1992). Theories of chess skill. *Psychological Research, 54,* 10-16.

Kasparov G. (2010). The chess master and the computer. *The New York Review of Books*; 57(2): 16–19.

Krogius, N. (1983). *Psychologie im Schach*. Berlin: Sportverlag.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences,* 12(7), 265-272.

Saariluoma, P. (1995). *Chess players' thinking: A cognitive psychological approach*. London: Routledge.

Samuel, A. (1959). Some studies in machine learning using the game of checkers, *IBM J. Res. Develop.* 3, 210-229.

Shannon, C. E. (1950). Programming a computer for playing chess. *Philosophical Magazine*, 41, 256-275.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning. An Introduction*. Cambridge, MA, MIT Press.

Tesauro, G.J. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6/2, 215-219.

Thrun, S. (1995). Learning to play the game of chess. In Tesauro, G., Touretzky, D. S., & Leen, T. K. (Eds.), *Advances in Neural Information Processing Systems 7* Cambridge, MA. The MIT Press.

Von Neumann and Morgenstern, 1944, *Theory of Games and Economic Behavior*. Princeton, Princeton University Press.

Appendix: Bronstein's Dream Game

White: David Bronstein

Black: David Bronstein

1. d4 Nf6 2. c4 e6 3. Nc3 Bb4 4. Bg5 h6 5. Bh4 Qe7 6. Nf3 d6 7. Qa4+ Nc6 8. d5!? ed5: 9. cd5: Qe4! 10. Nd2 Qh4: 11. dc6: 0-0 12. a3? Ng4 13. g3 Qf6 14. ab4:? Qf2:+ 15. Kd1 Ne3+ 16. Kc1 b5! 17. Qb3 Be6 18. Qa3 Qe1+ 19. Nd1 Qd1:+ checkmate.

0-1

Matteo Colombo
Tilburg Center for Logic, General Ethics and Philosophy of Science
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
m.colombo@uvt.nl

Jan Sprenger
Tilburg Center for Logic, General Ethics and Philosophy of Science
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
j.sprenger@uvt.nl