Preferential Engagement and What Can We Learn From Online Chess?

Vadim Kulikov[*◇]

November 3, 2020

An online game of chess against a human opponent appears to be indistinguishable from a game against a machine: both happen on the screen. Yet, people prefer to play chess against other people despite the fact that machines surpass people in skill. When the philosophers of 1970's and 1980's argued that computers will never surpass us in chess, perhaps their intuitions were rather saying "Computers will never be favored as opponents"? In this paper we analyse through the introduced concepts of psychological affordances and psychological interplay, what are the mechanisms that make a human-human (HH) interaction more meaningful than a human-computer (HC) interaction. We claim that an HH chess game consists of two intertwined, but independent simultaneous games – only one of which is retained in the HC game. To help with the analysis we introduce the thought experiment of a *Preferential Engagement Test* (PET) which is inspired by, but non-equivalent to, the Standard Turing Test. We also explore how the PET can illuminate, and be illuminated by, various philosophies of mind reading: Theory Theory, Simulation Theory and Mind Minding. We propose that our analysis along with the concept of PET could illuminate in a new way the conditions and challenges a machine (or its designers) must face before it can replace humans in a given occupation.

Acknowledgments This research has been partly supported by the PERCEPT project (322637) funded by the Academy of Finland and the HUMOR project (3656/31/2019) funded by Business Finland and industry.

1 Introduction

What are the conditions that a machine, or an artificial system must satisfy to replace humans in a particular occupation? Will computers ever replace humans in art, composing music, being mathematicians, or performing as rockstars? The general statement we want to defend in this paper is that there are situations in which computers *cannot* replace humans at all. This seems to be acute, because many thinkers today argue that the contrary is true. We do not claim that our defence is absolute, but we believe to have a novel point. This point might have other implications to philosophy of machine intelligence to be explored elsewhere. This novel point can be summarized as follows.

[*] Senior Research Fellow at the University of Oulu, Faculty of Information Technology and Electrical Engineering, Center for Ubiquitous Computing, Oulu, Finland

[◇] Adjunct Professor at the University of Helsinki, Department of Mathematics and Statistics, Helsinki, Finland

Suppose a human H interacts with entity X. The subjective situatedness of H depends on H's knowledge of whether X is a human or not. And if H knows that X is a computer, this makes certain affordances available to H which in turn enables H to exercise certain fine tuned skills which would not otherwise be used and remain idle. We will also show that according to some accounts of agency, the exercise of these fine tuned skills is tantamount to the sense of agency of H and therefore interaction with a human X (and knowing it) will be subjectively more meaningful than with a non-human X. To fully develop the argument we will define notions of *psychological affordances* and *psychological interplay.*

To illustrate the point we will focus on online chess as a test case. Online chess is a good example of a situation where computers have long outperformed humans, but we still prefer to play chess against each other, as well as watch games played by other humans against each other. In professional and amateur chess alike, the "engine" has merely the role of an encyclopedia and not of an interesting opponent. (Data collected from 20 randomly picked active members of the chess server `lichess.org` showed that the average total number of human-human games of these players was 3662.15, and for human-computer games it was 41.15 ($p < 0.0002$).)

To create a framework for this discussion, we introduce the *Preferential Engagement Test* (PET). This test is inspired by, but not equivalent to, the Standard Turing Test (STT). We will show that in some scenarios PET can provide relevant information on whether or not a machine can replace a human in a given occupation, information on which STT is silent.

In the early years of cybernetics and computer science scientists were divided: some predicted that one day computers will outperform humans in chess, others said that it will not be possible. See the Introduction in Dreyfus (1979) for a history of chess A.I. in 1950–1970. But now that computers have long outperformed humans in chess, we are witnessing a curious phenomenon: not only do people continue to play chess, but they do so against each other and not against the machines. We conjecture that the intuition of those philosophers and scientists who argued against the success of computers in chess was partially right. They argued that computers will never be better than humans in chess, but perhaps part of their actual intuition was that computers will never *replace* human players, and this is still the case.

As data presented above shows, chess engines fail the PET and we ask: Why? The answer we give is that human-human (HH) chess consists of two intertwined but still quite independent components, only one of which is retained in human-computer (HC) chess. In particular there are skill sets possessed by human chess players that are put into use in HH chess but are bound to be idle in HC chess making the HH game experience richer, more satisfactory, and more meaningful.

2 Preferential Engagement Test

Imagine the setup of the Standard Turing Test (STT). Player C (the interrogator) walks into a room with a pair of computer terminals. On one of the terminals Player C can exchange information with Player A and on the other one with Player B. Players B and C are both humans while Player A is a machine. Unlike in the STT, the interrogator *knows* which one is the human and which one is the machine. Player C now takes turns engaging with A and B, and after a while has to answer the question:

> (Q_{PET}) "Which one, A or B, would you like to continue to engage with for the next period of time T?"

If Player C answers "A", then the machine *has passed the Preferential Engagement Test* (PET). Another version of this test, called *Meaningful Engagement Test* (MET) is one where instead of (Q_{PET}), the interrogator is asked

(Q_{MET}) "Which interaction, the one with A or the one with B was more meaningful?"

We will focus exclusively on PET in this essay, because the concept of *meaningfulness* breads too much philosophical questions, confusions, and vagueness. However, the reader should keep in mind that MET would in ideal circumstances capture better what we have in mind when we define PET as a more tractable alternative. In light of the discussion of the relationship to the sense of agency (end of Section 7) one could hope connect these dots, but this is left for future work.

The assumption that the identities of A and B are disclosed to Player C in the beginning of the test is important. It will be argued in Section 7 that the mere knowledge that the other player is a computer (or human) might change the dynamics of the game as well as the way the game is experienced. A computer might fake human behavior perfectly, but once it is *known* that it is a computer, the human may lose interest in further engaging with it.

There are four theoretical possibilities for a given machine M. It can fail both the STT and the PET (Fail-Fail), it can fail the STT and pass the PET (Fail-Pass), it can pass the STT and fail the PET (Pass-Fail), and it can pass both the STT and the PET (Pass-Pass). Let us take a look at these four possibilities. Can we come up with natural examples and what are the implications?

Fail-Fail. Imagine that Player A is a poorly designed bot that not only fails the STT, but is also a boring conversation partner. For example maybe it only outputs the letter "z" all the time. Then, assuming that Players B and C get along, A will fail both the STT and the PET. This bot is not sophisticated, however. For a different example consider a chess engine. Skilled chess players can, at least for some algorithms, distinguish human game from computer game, whence computer fails the STT. From empirical data (see above) we know that humans tend to prefer to play chess against other humans.

Fail-Pass. Suppose Player B is impolite, offensive, and boring while Player A is Wikipedia. It will be easy for Player C to determine which one is human, but she might prefer the "interaction" with A. There is of course a problem with this scenario, because most of Wikipedia is written by humans. A more fair example could be found again from the realm of games. If Player B is a very weak opponent, while Player A is challenging and interesting, one might imagine a situation where the interaction with A is experienced as more engaging, if only for the game's dynamics sake.

Pass-Fail Let us base a speculative example again on the game of chess. It is not hard to imagine that with modern machine learning techniques a chess bot could fake human behavior quite well. After all, there are not many variables to fake: time that it takes to make a move and the quality of the moves. Thus, it is probably quite easy for a chess bot to pass the the STT, especially if the interrogator is not a professional chess player. However, for the same reason as in Fail-Fail scenario, we expect the computer to often fail the PET. In this case the difference is only in *knowing* who your opponent is.

Pass-Pass To tap into the realm of science fiction, imagine a scenario similar to the one in Fail-Pass, but instead of Wikipedia there would be an exceptionally interesting, intelligent and mesmerizing artificial intelligence which is capable to pass the STT.

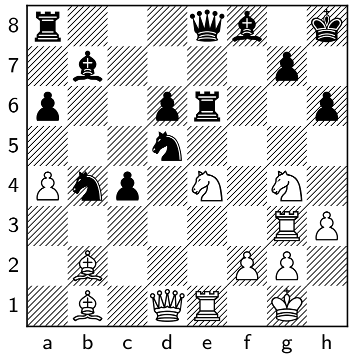


Figure 1: Kasparov-Karpov (1990) In this position white played **26 ♘×h6**. Black did not recapture, but played **26. . . c3** instead.

3 The Value of Psychological Interplay in Chess

Before the theoretical definition of psychological interplay theoretically in Section 4, let us look at some concrete examples – again from chess.

In a world championship game Kasparov-Karpov, Kasparov sacrificed his knight for a pawn on move 26, but Karpov did not accept the sacrifice.[1] In a videotaped discussion about this game the interviewer asked Kasparov "[Did Karpov reject the sacrifice] because he was confident in your calculations?" and then added "Or in your feelings?", and Kasparov answered without a second thought: "In my feelings." (Figure 1)

That is to say, according to Kasparov, the reason for Karpov not to take the knight was not (entirely) a matter of his own (Karpov's) calculation, his sophisticated understanding of the combinatorics of chess, not of his own feeling, and not even of his confidence in his opponent's (Kasparov's) calculations, but it was (at least partially) a matter of Karpov's confidence *in Kasparov's feelings*. And note that this is the analysis supplied to us by Kasparov, not Karpov. This is a perfect example of psychological interplay in advanced chess. Whether this analysis by Kasparov of Karpov's decision making was correct or not is irrelevant for the present discussion. What is relevant is that it gives us a clue towards the kind of psychological engagement that can happen in chess. If Kasparov's post-hoc analysis was that the actual reason for Karpov not to take the piece was based on attributing certain psychological states to Kasparov, then who knows what was going on in his mind during the actual game.

Another example is the following quote, also by Garry Kasparov describing another position in the same tournament:

> "I managed to find a way of maintaining the opponent's feeling of discord."
> (Kasparov, 2010, p. 322)

How much do such psychological attributions, attributions of attributions, and so on, play a role in decision making in chess? Are they being conceptualized during the game or only in aftermath? This would be a fascinating topic for empirical research on its own.

4 Psychological Affordances and Interplay: Theoretical Definition

One can think of social interactions as a game where each action performed by the agents is their "move". In regular face-to-face interactions the moves are spoken words, body language,

[1] G.Kasparov - A.Karpov World Championship Match 20th Game, Lyon 15.12.1990 Ruy Lopez C92

gazes, gestures etc. This framework of looking at social interaction as a game is agnostic towards the philosophical position concerning theory of mind. Let us further examine the metaphor of a game. In our definition of a game we demand to conditions to be satisfied:

(G1) follow the rules and

(G2) be able to predicate their behavior on the assumption that the other participants follow the rules.

The idea is that once (G1) or (G2) is violated, the game ceases to exist.

In essence (G2) is usually captured by "knowing that my opponent is also playing the same game as me". We will use this in Section 5 to show why the psychological interplay in chess exists as long as both players *know* that the other is a human.

If during a soccer game a player notices that other participants start touching the ball with their hands, she may lose her trust in the fact that the soccer game is being played anymore. A professional soccer player's skill set is useful and can be exercised only as long as the other players are following the rules of the game, otherwise the player's skills become obsolete. When everyone is following the rules the player may predicate her actions on the following assumptions in a descending order of certainty:

(a) the other players have constraints ordained by the commonly accepted rules,

(b) the other players predicate their actions on the assumption that (a),

(c) the other players predicate their actions on the assumption that (b),

⋮ ⋮ ⋮ ⋮

Some of those initial constraints are explicit (like "it is forbidden to touch the ball with hands"), some are implicit. For example the rule that a player cannot teleport from one side of the field to the other is an implicit constraint. It is not an explicit rule of the game, but everyone's apparent behavior is predicated on the assumption (even if not explicit or represented) that everyone is subordinate to the non-teleportation constraint. This constraint is a consequence of the type of embodiment the players have. In a virtual soccer game played on a computer screen such moves are theoretically possible. This is why it is not enough for a computer game designer to be aware of the rules of soccer, but also of the relevant laws of physics and human anatomy.

By our analogy, every social interaction, if it can be viewed as a game, has a set of rules and constraints. The participants who are engaged in the interaction must be playing the game in the sense of (G1) and (G2) above. We are still being agnostic on the question whether these rules and constraints are represented in the participant's minds or not, and what is the mind reading paradigm involved. We simply introduce a metaphor to talk about social interactions.

4.1 Affordances predicated on the constraints.

A participant in a game has certain affordances that comply with the rules of the game and are predicated on the assumptions (a), (b), (c), etc. For example in soccer, a player may have an affordance of making a goal predicated on the assumption that the goalkeeper does not use devices that are forbidden by the rules of the game. In a soccer game, a player has the affordance to run freely when there are no other players around, but only predicated on the assumption that the other players will not suddenly teleport from all corners of the field to surround her. Further, there are affordances that are predicated on the fact that the other player's affordances are predicated on each other's constraints and so on.

In other social interactions we can talk about affordances in the same way. In a face-to-face conversation one can be considered to have the affordance to evoke certain emotions in the other participant by making certain facial expressions, gestures or by changing ones own voice. In an acrobatics class Person A can have an affordance of standing on Person B's shoulders, but only predicated on the assumption that B has the relevant skill and desire to collaborate. In a mathematics class room there is (usually) no socially acceptable affordance to start riding a horse. If someone appears to the mathematics class riding a horse, the class will be interrupted and the "social game" ceases to exist; the participants are no longer following the necessary constraints of a mathematics lesson or can no longer trust that others follow them.

4.2 Three Parallel Accounts

It is hard to continue the discussion of psychological affordances and interplay without fixing a philosophical framework of mind reading. In order to show, however, that psychological affordances and interplay are concepts that make sense in various such philosophies, three different accounts will be presented corresponding to Theory Theory, Simulation Theory and Mind Minding accounts of mind reading. The main focus is on Mind Minding, however.

4.2.1 Theory Theory Account

Theory theory (TT) is the position that people understand social situations through inference about others inner mental states. Gallagher (2008) describes it as follows:

> Proponents of [TT] contend that inference formation happens as the result of a mental consultation with a theory or a set of folk-psychological rules that will allow one to deduce an explanation of the observed behavior in terms of beliefs and desires understood as the other's mental states.
> (Gallagher, 2008)

Folk-psychological rules are all void, if the target is not a human. If we are used to making such inferences with respect to humans, taking into account human biases and human type of reasoning, then it should be hard, if not impossible, to make similar inferences when the target is non-human. Note that the assumption that the target is a human implies that the target is also engaged in a TT-inference and this should be taken into account by the attributor too. This seemingly leads either to circularity or infinite regress, but Friston and Frith (2015, p. 130) argue that "this infinite regress dissolves if the two brains are formally similar and each brain models the sensations caused by itself and the other as being generated in the same way." Thus, either way, such engagement, or "theorizing" becomes impossible if the involved brains are vastly dissimilar to each other.

The examples of Section 6 should be illustrative to a proponent of TT of why a big chunk of our mental inferential skill set is left unexploited when playing against a computer (or other type of an alien mind).

4.2.2 Simulation Theory Account

Simulation theory (ST) is an embodiment and enactivist friendly account on mind reading. It holds that we use our own mental processes to simulate another persons intentions, beliefs and desires. Here is an example of a statement how it could work:

> First, the attributor creates in herself pretend states intended to match those of the target. ... The second step is to feed these initial pretend states into some mechanism of the attributor's own psychology ... and allow that mechanism to operate on the

> pretend states so as to generate one or more new states. Third, the attributor assigns the output state to the target.
> (Goldman, 2005, pp. 80-81)

It is easy to imagine how this could work in chess. If I see my opponent making a certain move, I imagine making such move myself. I then let my normal brain processes arrive associatively (or through "simulation" whatever that means in this context) at a possible reason or idea which I could have in mind if I was making such a move. This is completely analogous to a situation in which I see someone moving their hand in a particular way, imagining (e.g. through mirror neuron activation) that I am making a similar move and then, through simulation, arriving at an understanding of the possible goals and reasons of such movement. Obviously, my brain will be led astray by such simulation, if the target is not a human. Presumably, if I *know* that my opponent is not a human, then either my brain would not go to the trouble of doing such simulations in the first place, or it would make them and arrive at mistaken conclusions which would result in non-anticipated patterns from which it is hard (for a human) to learn and in general displeasure and cognitive dissonance.

Note that, as Gallagher (2008) puts it, both theories, ST and TT, share the fundamental assumption that "the problem is best posed as one that involves lack of access to other minds." Despite this limitation, it is clear that certain skills that we have of "mind reading" are left unused when we engage in an interaction with a non-human mind.

4.2.3 Mind Minding Account

Mind Minding (MM) is a radical enactivist account of mind reading (Hutto, 2011, 2017). According to radical enactivist view on cognition, social engagement with other people is a skill and a type of activity which has more to it than mere inference and simulation. According to the Mind Minding Hypothesis,

> [B]asic acts of social cognition involve being responsive to the intentional attitudes of others in ways that do not involve representing or attributing any kind of mental state concepts or contents.
> (Hutto, 2017)

Gallagher (2008) writes about the enactive view on social cognition:

> When I enter a classroom or a grocery store, I can immediately see who the teacher is or who the cashier is, and I can intuitively understand what they are doing, and for my particular purposes that may be sufficient for my interactions.
> (Gallagher, 2008)

In this view, the context of a situation together with previous experiences and cultural norms determine social and psychological affordances. What kind of behavior is expected of the agent and what are its readiness to execute this or that behavior is implicit, rather than explicit, in the way the agent's brain and body embed into the social and cultural context.

Here is passage from Gallagher (2008) which contrasts TT and ST with the enactivist notion:

> Rather than making an inference to what the other person is intending by starting with bodily movements, and moving from there to the level of mental events, we see actions as meaningful in the context of the physical and intersubjective environment. If, in the vicinity of a locked door, I see you reach for a set of keys, I would know your intentions as much from the door and the keys, your bodily posture and expression as from anything that I postulate in your mind. We interpret the actions of others in terms of their goals and intentions set in contextualized situations, rather than

> abstractly in terms of either their muscular performance or their beliefs. (Gallagher, 2008)

The above will only work, if all participants are actually human and none of them is a robot in disguise. Recall that in the PET the identities of Players A and B are revealed to the interrogator and so even if the robot is able to perfectly fake human behavior, the mere knowledge that it is not human might be enough to destroy the psychological interplay (see Section 5).

In a situation where two or more humans are present, are able to observe each other's bodies and hear each other's voices, the social affordances may be quickly determined. Social affordances, for example, include facial expressions which have certain meaning and emotional consequences. Change of intonation and other subtle embodied cues constitute social affordances which enable two or more people to be coupled in a social interaction forming a system which is, the enactivists hold, better analyzed as a whole, and not as two or more brains trying to make inferences about each other. But what happens, if we constrain the communication channel? What about a conversation over the phone? What about a chat? What about an online chess game?

The enactivist take is that when one finds oneself in the situation of playing an online chess game against another human, one experiences certain psychological affordances (see Section 4.1). There is phenomenological "readiness" to get socially coupled to the opponent even though the communication channel is narrow. Note that the *minimalist perceptual crossing paradigm* has supplied a lot of empirical evidence that social cognition does not require channels of high bandwidth (Auvray, Lenay, & Stewart, 2009; Lenay, 2012; Deschamps, Lenay, Rovira, Le Bihan, & Aubert, 2016).

5 Psychological Interplay and the PET

What is the connection between psychological interplay defined in Section 4 and the Preferential Engagement Test? As we saw, when the minds engaged in a social activity are similar and familiar to each other, then these minds can exercise skills that they otherwise cannot put to use. Humans are extremely skilled at detecting and adapting their own behavior to cues such as gaze direction, attentional targets, intentions, gestures, body language, vocal intonation etc. This adaptation is predicated on the social situation and various constraints that are present. Thus, people form a coupled system where each move and gesture is predicated on the assumption that the other agent is a human too. But if the other agent is not human, then a wide range of all these cues lose their meaning. In terminology of Section 4, the player can no longer predicate its actions and affordances on (a), and even if she can, she cannot predicate them on (b), and even if she can she cannot predicate them on (c) and so on. Thus, even if Player A (the machine) passes the STT, once Player C *knows* that A is a computer, C is no longer necessarily able to trust that A will engage in a meaningful psychological interplay with C. The psychological game ceases to exist. The Player still trusts that the opponent follows the rules of chess, so the chess game continues to exist. When Player C knows that A is human, she can rely on certain affordances that are provided by a human opponent, she can rely on the fact her opponent is also relying on similar affordances and so on, i.e. clauses (a), (b), (c)... of Section 4 are satisfied. Revealing that the opponent is a computer is analogous to telling the soccer player that all the other players actually posses the ability to teleport, they just didn't use it up until this point out of politeness. Then the soccer player can no longer rely on the typical affordances of a soccer game and cannot exercise his skill set in a normal way.

6 Chess Engines and Preferential Engagement

We can now argue, based on Sections 3, 4 and 5, that in HH chess there are two games being played simultaneously. First, the actual combinatorial game defined by the rules of chess. Second, a game based on a psychological interplay between the players. The rules of that latter games are mostly implicit and non-representational in nature. That game exists only insofar as the players can predicate their behavior on the implicit or explicit assumption that the other player has certain constraints and that the other player is (implicitly or explicitly) aware of that predicament.

In HC game only the combinatorial game is retained along with a possible illusion of the psychological interplay. This is why many skills that a human chess player has are left unused in a HC game and only part of the rich experience of a HH game is enjoyed.

The difference between HH and HC chess, involving the value of the psychological interplay, is most vivid under the radical enactivist account (Section 4.2.3). This is because under this account the psychological coupling of two agents which is the result of the right type of readiness to engage in the coupling exists in a HH game, but not in a HC game. According to all three accounts, Theory Theory, Simulation Theory and Mind Minding, there are unused skill sets a HC game compared to a HH game. In the case of TT these are inferential skills, in the case of ST these are the capabilities of simulating the other player's mental processes and in MM they are anticipatory mechanisms, intentions, complex cue reading and response abilities. But only MM claims that an important part of the formation of a dynamical system through coupling, as if a third entity is created.

I will now go through several examples and illustrations of what kind of psychological affordances can be present in HH chess. Each of those examples is speculative and anecdotal in nature, but gives predictions and suggestions for potential empirical research.

6.1 Sunk Cost Fallacy

In psychological and economic literature on decision making, *the sunk cost effect* or the *sunk cost fallacy* is a phenomenon in which an agent makes an economically irrational decision because of prior incurred irreversible costs. For example a person who bought a ticket to an event in advance, but is not enjoying the event, might be less likely to walk away than a person who got to the ticket for free. There are several possible psychological factors at play such as loss-aversion, social acceptance, sense of responsibility etc. (e.g. Arkes, 1985).

Sunk cost effect can manifest itself in chess. A simple example is when a piece has been sacrificed in order to execute an attack on the opponent's king. Even if the attack fails due to an unexpected move by the opponent, the the attacker may feel the urge to still try to continue or resurrect the attack. This can lead to moves (for better or worse) that would not otherwise been considered. It is conceivable (but ultimately an empirical question) that chess players are not only implicitly aware of their own sunk cost fallacy, but at times appear to rely on similar tendencies of the other player.

6.2 Minimization of Expected Cognitive Load

A human might prefer to make moves which (according to subjective judgement) result in a position that is cognitively easier to process. The opponent, on the other hand, may understand this desire and choose her moves accordingly. If we take the clock into account, one can argue that humans may challenge each other with moves that take longer (for a human) to decide upon. Here, as in other examples, it is important to note that this is not (necessarily) done through an explicit or conscious effort. On the view presented in Section 4.2.2 it could be an integral part of psychological engagement, an enactive mind minding interplay.

It is well known that in a position where one side has only the king and the other side has a king and a queen, most amateur chess players follow one of the predetermined “checkmating strategies” which can take up to ten times more moves than necessary to achieve the goal, even though with little cognitive effort they would surely find a faster way.

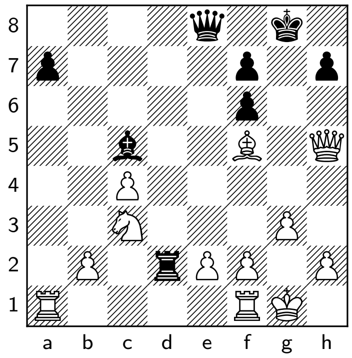
6.3 Commitment to Plans and Affordance Bias

The dynamical systems account of action planning and execution (Smith, Thelen, Titzer, & McLin, 1999; Thelen, Schöner, Scheier, & Smith, 2001) suggests that an action that has been planned for is more likely to be executed. Imagine that you are about to move your queen, but, unexpectedly, your opponent makes a move which makes moving your queen a bad idea. Can it be that moving your queen is now more likely than if you were not planning to move it in the first place? Even if it is not, can it be that your next move is still *somehow* influenced by the prior, but now discarded plan? Do human players (subconsciously) use this phenomenon to divert their opponent’s attention?

In one of his teaching videos Grandmaster Benjamin P. Finegold says that beginner chess players are likely to capture a piece just because it is possible to do so (Finegold, 2020). From a cognitive perspective this observation seems to fit the idea that perceived affordances are likely to be acted upon even when other, more hidden, but more valuable opportunities are available (even consciously to the acting agent).

6.4 Spatial bias

The spatial location of the pieces may invoke false judgement in humans. For example a piece being spatially closer to the king can be perceived as defending it. Here is an example.



In this position black made the move **. . . ♕f8**. It is the worst move in this position for black, because after this there is one-move checkmate for white **♕×h7**.

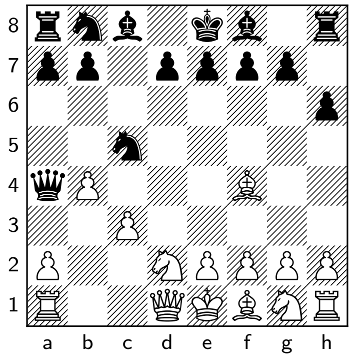
The phenomenological account of the player who was playing black pieces (the author) is: “Under the time pressure I had to make a quick decision. Since the king was threatened, I thought that it needs to be protected, so I moved the queen closer to the king because it felt like it will protect it.”

This is illustrative of what kind of spatial intuitions maybe guiding our psychological game. Not only the player of black pieces is susceptible to making a move based on spatial intuition, the opponent too could interpret the move as an attempt to protect the black king, because the opponent was also human and had similar biases. In some circumstances (unfortunately not in the game just described) the opponent could even slow down the attack if he or she took seriously black player’s intuition and (mistakenly) thought that the move *actually* protected the black king. Or perhaps, on the contrary, the white player consciously or subconsciously *counted* on the black player making such a mistake.

This is an example of a possible psychological battle between players which is possible only if two conditions are satisfied: both have spatial bias and both expect of each other that they have (some form of) spatial bias corresponding to (G1) and (G2) of Section 4.

6.5 Attention Bias

Human player's attention can be drawn to a particular part of the board given certain cues. This might distract the player from another part of the board which also demands attention. This, again, can be exploited by skilled players in each other. Also the tendency to exploit such attention bias can be a bias in its own right etc. Here is an example:



What is the best move for white? The attention of the human player maybe drawn to the area **a4**, **b4**, **c5**. If white captures **b×♘5**, then black captures **. . . ♕×f4**; this is not the best move for white. The best move is **♗×b8**, because now e.g. after recapturing **. . . ♖×b8** black loses the knight on **c5** since the bishop is no longer on **f4** to be captured by the queen. The move **♗×b8** is counter-intuitive, because it is not a typical move in an opening to exchange an active bishop for a passive knight.

6.6 Conclusion: Two Facets of Human Chess

The game of chess between two humans has two components. One is the actual combinatorial game chess and the other is the psychological game. When two people engage in a game of chess they are simultaneously playing these two games. But when a person is playing against a non-human opponent, only one of the games is being played. In terms of System 1 and System 2 (Kahneman, 2011; Stanovich & West, 2000), one could roughly characterize these two games as played by System 1 (the psychological interplay) and System 2 (the combinatorial game). My hypothesis is that the psychological game is an indispensable part of a rich experience when it comes to playing chess, and is one of the reasons for why people prefer a human opponent over a machine.

The rules of chess are clear and unambiguous and satisfy conditions (G1) and (G2) of Section 4. For successful game both players must adhere to them and they both have to trust that the other player adheres to them too for otherwise the game loses its meaning.

The rules of the psychological interplay are less clear and more ambiguous, but they do constrain both players in specific human ways. As long as both are aware that the opponent is also human, conditions (G1) and (G2) are satisfied also for the psychological interplay.

7 Preferential Engagement Test vs. Turing Test

One of the main aims of the originally formulated Turing Test, initially termed the Imitation Game (Turing, 1950), was to reframe the *hard* question "Can machines think?" in a more rigorous and empirically tractable manner.

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly ... ? These questions replace our original, 'Can machines think?'
> (Turing, 1950)

Thus, Turing suggested that a positive answer to "Can machines imitate humans?" would be predictive of a positive answer to "Can machines think?"

Famously J. Searle proposed another thought experiment known as the *Chinese Room Argument*:

> Searle imagines himself alone in a room following a computer program for responding to Chinese characters slipped under the door. Searle understands nothing of Chinese, and yet, by following the program for manipulating symbols and numerals just as a computer does, he sends appropriate strings of Chinese characters back out under the door, and this leads those outside to mistakenly suppose there is a Chinese speaker in the room.
> (Cole, 2020)

One of the conclusions of this argument is that "to appear to understand a language" is not the same as "to actually understand a language". Here "understand a language" can be replaced by "think", "be conscious", "be aware", "have feelings" etc., corresponding to almost any type of *hard* question, in particular questions about consciousness and self-awareness. In this way Searle argues that a positive answer to Imitation *is not* predictive of, and certainly not a guarantee for, a positive answer to Thinking, Consciousness and Awareness.

Searle's argument was adapted by embodied mind theorists who use it to argue for the Symbol Grounding paradigm, namely that symbols cannot be meaningful unless their meaning is grounded in sensorimotor interactions with the world (e.g. Barsalou, 1999). By another argument of Harnad (1990) symbols' meaning cannot be grounded in other symbols either. Harnad (1990) compares the situation with someone who does not know any Chinese trying to learn the language from a Chinese-Chinese dictionary.

Thus, it would seem, that the same group of arguments which can be used to argue against the validity of the Turing Test can be used to argue that a device which only manipulates symbols can never have any semantic understanding of the meaning of those symbols, i.e. the device which bears these symbols does not bear their contents.

Can the PET do any better and survive such objections? We claim that under the philosophies TT and ST (Sections 4.2.1,4.2.2), or any other representations involving philosophy, the answer is "no", but the radical enactivist account (Section 4.2.3) has the potential for "yes".

So suppose that Player A is a bot which consistently passes the PET. In light of the theory presented in this paper this might be due to the following reason. Not only does the interrogator learn – during the first phase of the experiment – which are the "psychological constraints" of the bot, but is also convinced that they are not fake (in the sense that they cannot be removed by a decision of Player A to do so). Moreover the interrogator feels that Player A responds in interesting ways to *her* psychological constraints and uncovers psychological affordances that the responses of Player A allow for, and so on, so that a non-trivial and engaging coupling between Player A and C is possible. This coupling is so successful that the interrogator prefers it over the one formed with Player B. On the TT account it is hard to see how this would imply anything about the inner states or contents in Player A's "head". If such psychological interplay is merely a result of inferences, these inferences can be successfully carried out as symbol manipulations (even if they need to be very sophisticated ones). These symbols need not carry any meaning in the beginning nor in the end. The apparent behavior is still the same. The story with the ST is similar. Simulations need to be *of something*. Why would the fact that the coupling is successful change the metaphysical essence of a given process from "mere stream of zeros and ones" into a "simulation of something in the outside world"? See also (Hutto & Myin, 2012, 2017) for the analysis of why even most enactivist and embodied theories of cognition are content involving.

Radical Enactivist view on Cognition (REC) is not subject that problem, because REC rejects the involvement of content in basic cognition anyway. Thus the problem is not there to begin with. However, since Player A was preferred by the interrogator, that is an indication of the fact that the bot is capable of complex and nuanced patterns of responses to cues, producing cues to which the interrogator responds and so on in a dynamic fashion. The mere capability and readiness of such coupling, according to REC, is already constitutive of some form of cognitive function on the part of Player A. However, REC is not purely a behaviorist theory and presupposes also some forms of intentionality. According to Gallagher (2017)

> [Enactive i]ntentionality is determined by what the agent is doing and what the agent is ready to do, and is constrained, for example, by the agent's sensorimotor skills relevant to coping with the situation at hand, whether that's stepping off a curb or stepping on the brake, or any interaction that might follow.
> (p. 79 Gallagher, 2017)

In our context we simply replace "sensorimotor skills" by "interactive skills" through the information channel available to Player A. Now things like "skill", "readiness to do something" and "constraints" are things that are on the one hand intrinsic to the organism's design, but on the other hand are subject to probing and testing by other agents, e.g. the interrogator. So if we were to trust Player C on his or her ability to judge the skills and strategies of Player A, then we should trust that since Player C experienced meaningful coupling with A, the latter ought to have those skills and strategies and therefore also intentionality on the enactivist account. Even though the above inference contains a couple of leaps of faith, it arguably comes closer to answering the *hard* questions (Q5) and (Q6) than the representationalist rivals.

Sense of Agency The kind of specific psychological interplay that is according to our analysis possible only in a HH game and not in an HC game is sometimes argued to be constitutive of the sense of agency.

> There must be a distinction available between changes in her body and in the environment that are due to her own agency and those changes that are due to other factors in the environment or her sensorimotor system.
> (Hohwy, 2007)

If we replace bodily interaction with more abstract interaction of the chess game, we may find that the skill that a human chess player possesses in understanding the human opponent will enable her to better distinguish between the types of moves that are "reactions" to her moves and other moves that originate from the opponents' own considerations. In this way in the HH game the player's own sense of agency could be enhanced.

8 Alternative explanations?

We explained the preferential engagement in HH chess over HC chess by invoking the notion of psychological interplay (Section 4) and arguing that the HH game is objectively *richer* and engages the participants on a wider scope of skill sets than the HC game.

There, of course, can be other explanations of the phenomenon. For example one might argue that it is impossible to win when playing against a computer, so it makes the game boring. But the computer's level can be easily adjusted to match one's own, so this explanation is untenable.

Another candidate explanation goes as follows. Establishing species-intrinsic hierarchies is presumably an integral part of mammalian biological drive. This is why humans want to compete and establish hierarchies between each other, not against the machines. This explanation might be partially correct. However, it is insufficient for the following reason. Instead of playing chess against each other people could all just play against the machine and gain points in this way. These points would then determine the hierarchy. This would be, in fact, a cleaner and a more controlled way to determine who is better at chess.[2]

At the end of the day the point of this paper is not so much to argue for the correctness of the psychological interplay explanation for chess (although I do believe that it is correct), but rather to give a theoretical framework for a wider range of phenomena captured by preferential engagement and social cognition.

9 Conclusion

We presented a new type of test, the Preferential Engagement Test (PET), which is inspired by, but not equivalent to, the Standard Turing Test (STT). The idea was to uncover new ways of approaching questions about machine intelligence, machine consciousness and the replacement of humans by machines. What we are secretly "really" after is *meaningful* engagement, but for the sake of philosophical clarity we analysed the concept of preferential engagement. We used chess as our main example. Chess algorithms seem to fail the PET despite the fact that they are vastly superior to humans in chess already for decades. This led us to explore the notions of psychological affordances and psychological interplay as indispensable parts of human-human (HH) interaction which are extremely difficult to replicate in a human-computer (HC) interaction: no matter the skill of the machine, the mere knowledge by the human that the other player is a machine might "break the spell". This is because at that point, the human cannot any longer rely on the other player to have the known and familiar psychological constraints as well as affordances (even if they have been faked up until that point by the computer). This immediately deprives the human of those affordances that would have been predicated on those assumptions concerning the other agent and so the psychological game "ceases to exist" in the same way as the game of soccer ceases to exist if all players acquire an unprecedented skill (like teleportation). Such skill has not been explicitly forbidden by the rules of soccer, but has always been an implicit part of the game. We go as far as arguing that according to some accounts of the sense of agency, the HC game also deprives the human player of that.

We give three accounts of psychological interplay corresponding to three philosophies of mind reading: Theory Theory, Simulation Theory and Mind Minding. We come to the conclusion that the latter is the most informative and provides the most vivid explanation of why the game of chess indeed is (objectively) more richer when both players are humans than when one is a computer. Hence the HH game results in a more fulfilling experience subjectively too, especially because it engages certain human skills that are otherwise idle. We support this point by giving examples of human-type psychological biases in chess. Such biases supply players with conditional (predicated) affordances – affordances that exist only *predicated* on certain assumptions about the opponent's behavior.

[2] Power lifting is an example of a sport where people get "points" by lifting various weights and then compare those points to establish a hierarchy. Of course power lifting can be done by machines, but it is trivial to argue that if a person outsources his power lifting to a machine, then that will diminish the richness of the experience (because there will be no experience).

Finally, in Section 7 we argue that the PET can withstand some of those objections that are posed to the STT by the Chinese Room type of arguments. However, this is the case only when the underlying philosophy of mind is aligning with radical enactivist views on cognition. In this way, the newly introduced concept of Preferential Engagement is shown to have relevance in explaining psychological phenomena and for philosophy of mind in general.

Needless to say that when analyzing political and philosophical questions related to Artificial Intelligence such as whether machines might or not replace humans in certain occupations and social roles, the PET is also relevant. It might in some circumstances provide a more adequate "intuition pump" than either the question of performance or the STT. Even if a machine can fake or supersede human skill, humans may still want to continue to engage with other humans for the reasons presented in this article.

References

Arkes, C., Hal R. nd Blumer. (1985). The psychology of sunk cost. *Organizational behavior and human decision processes*, *35*(1), 124–140.

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New ideas in psychology*, *27*(1), 32–47.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav Brain Sci.*, *22*(4), 577-609.

Cole, D. (2020). The chinese room argument. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/chinese-room/.

Deschamps, L., Lenay, C., Rovira, K., Le Bihan, G., & Aubert, D. (2016). Joint perception of a shared object: A minimalist perceptual crossing experiment. *Frontiers in psychology*, *7*, 1059.

Dreyfus, H. (1979). *What computers can't do: the limits of artificial intelligence.* Harper & Row.

Finegold, B. P. (2020). *Channel GMBenjaminFinegold.* YouTube LLC. Retrieved from https://www.youtube.com/channel/UC6EnFbK-P5q0zeaqI5yobKg

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *cortex*, *68*, 129–143.

Gallagher, S. (2008). Understanding others: embodied social cognition. In *Handbook of cognitive science* (pp. 437–452). Elsevier.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind.* Oxford University Press.

Goldman, A. (2005). Imitation, mind reading, and simulation. *Perspectives on imitation: From neuroscience to social science*, *2*, 79–93.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346. doi: 10.1016/0167-2789(90)90087-6

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche*, *13*(1), 1–20.

Hutto, D. D. (2011). Elementary mind minding, enactivist-style. *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*, 307.

Hutto, D. D. (2017). Basic social cognition without mindreading: minding minds without attributing contents. *Synthese*, *194*(3), 827–846.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content.* MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism.* MIT Press.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kasparov, G. (2010). *Garry kasparov on modern chess, part 4: Kasparov v karpov 1988-2009.* Gloucester Publisher.

Lenay, C. (2012). Minimalist approach to perceptual interactions. *Frontiers in Human Neuroscience*, *6*, 98.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: the task dynamics of the a-not-b error. *Psychological review*, *106*(2), 235.

Stanovich, K. E., & West, R. F. (2000). Individual difference in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-726.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and brain sciences*, *24*(1), 1–34.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.