Social Cognition in Silica
A 'Theory of Mind' for Socially Aware Artificial Minds

Michael Harré

Complex Systems Research Group, Faculty of Engineering and IT, The University of Sydney, Sydney, 2006, NSW, Australia

Keywords: Theory of Mind, Artificial Intelligence, Social Networks, Game Theory.

Abstract: Each of us has an incredibly large repertoire of behaviours from which to select from at any given time, and as our behavioural complexity grows so too does the possibility that we will misunderstand each other's actions. However, we have evolved a cognitive mechanism that allows us to understand another person's psychological space: their motivations, constraints, plans, goals and emotional state and it is called our 'Theory of Mind'. This capability allows us to understand the choices another person might make on the basis that the other person has their own 'internal world' that influences their choices in the same way as our own internal world influences our choices. Arguably, this is one of the most significant cognitive developments in human evolutionary history, along with our ability for long term adaptation to familiar situations and our ability to reason dynamically in completely novel situations. So the question arises: Can we implement the rudimentary foundations of a human-like Theory of Mind in an artificial mind such that it can dynamically adapt to the likely decisions of another mind (artificial or biological) by holding an internal representation of that other mind? This article argues that this is possible and that we already have much of the necessary theoretical foundations in order to begin the development process.

1 INTRODUCTION

One of the key drivers of work on the development of artificial human-like reasoning has focused on how we come to understand the inanimate world. For example, it is not uncommon to discuss cognitive development almost solely in terms of the cognitive structures relating to inanimate objects (see (Tenenbaum et al., 2011) for an example, where work on Theory of Mind is mentioned in *Open Questions*). While the learning of inanimate relationships is an important direction to explore, it is only one piece of the puzzle of cognition. In contrast, the focus of this work is on our ability to understand interpersonal *animate* relationships, an ability that appears to be equally old and important in evolutionary terms as other forms of comprehending the world.

Earlier work in anthropology has revealed some striking facts about our neuro-cognitive architecture and the role it plays in our social development as a species. In 1992 Robin Dunbar put forward the hypothesis that our neocortex grew in size in order to accommodate the cognitive pressures imposed upon us by our growing social group size. One of the notable predictions to come out of this work is that humans should have a social group size of approximately 150 individuals (Dunbar, 1992), this has been called Dunbar's Number. The relationship between the neocortex size and social group size is called the *Social Brain Hypothesis* (Dunbar and Shultz, 2007) and over the last 20 years considerable evidence across many primate and non-primate species has been collected supporting the hypothesis (Kudo and Dunbar, 2001; Shultz and Dunbar, 2007; Lehmann and Dunbar, 2009). Perhaps most significantly, recent studies have shown that this hypothesis is true at the individual as well as at the species level. In a study published last year (Powell et al., 2012) by Dunbar and colleagues it was shown that in individual humans, the size of their social network was linearly related to the neural volume of the orbital prefrontal cortex. A secondary but critical finding of the same study is that neural capacity is necessary but not sufficient for large social networks, the subjects also needed to have developed the psychological skills necessary for understanding another persons point of view. This cognitive skill, called *Theory of Mind* (ToM (Baron-Cohen et al., 2000)), enables us to understand that other people have mental states and that these states might be different from our own. This enables us to better manage our social relationships and maintain a larger, more complex and more diverse set of social relation-

Harré M..
Social Cognition in Silica - A 'Theory of Mind' for Socially Aware Artificial Minds.
DOI: 10.5220/0004917606570662
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 657-662
ISBN: 978-989-758-015-4

ships than any of our primate relatives.

Following on from these studies, the aim of this article is to introduce a simple model of social interactions between artificial agents that imply a view of the agent as filtered through the strategic perspective of another agent, this is called a *strategic Theory of Mind* (Harré, 2013). This is then extended to interactions of multiple agents in a social network such that a single agent's decision-making process is influenced by those in closest proximity in their social network, but these closest relationships are in turn influenced by second order relationships that are not directly related to the first agent. The result is an extension of strategic ToM to a socio-strategic ToM and social networks in general.

2 SPECIFIC FUNCTIONAL ROLES OF THEORY OF MIND

Theory of Mind research looks at how we are able to reason based on an internal representation of how an individual believes other people's minds operate in general and then to use this representation to understand how specific contexts influence another individual's actions. In strategic interactions, such as economic game theory, understanding another person's state of mind has the direct and obvious advantage of benefiting in terms of increased payoffs (Bhatt and Camerer, 2005), but these benefits extend to every aspect of our lives, to how we teach children, collaborate in scientific research, empathise with the less fortunate and how the economic division of labour allows us to divide tasks according to the specific skills and abilities of each individual.

In order to understand the neural foundations of our ToM, recent progress has been made in the neuroimaging of human subjects carrying out ToM related tasks. A complex network of brain regions have been revealed that are activated during any cognitive task that involves thinking of another person's state of mind or even social interactions with animate rather than inanimate agents. Focusing specifically on understanding and internally representing the mental states of others, two of the most important functional properties of these brain networks are the ability to recognise that people, unlike other things in the world, have mental states that include thoughts, feelings, constraints, goals and perceptions and the development of an internal model of how these mental states influence the decisions they make within a specific environmental context (Lieberman, 2007). In this article, the focus will be on person A thinking of the environmental context in which person B is making decisions, and B's environment will be a social environment (that may include A). Note that this is only a subset of the possible contexts in which B could be making a decision and A might still find it useful to have a ToM for B in such contexts, but this is not the focus of this article.

A special case worth highlighting is perspective taking, which most commonly refers to understanding the sensory perception of another person, for example that another person sees something different to what you can see. Humans can solve perceptual perspective-taking tasks using visuo-spatial reasoning without the need of a ToM mechanism (Zacks and Michelon, 2005). However there is a broader meaning to perspective taking that includes adopting or considering another person's *psychological* perspective, and this is sometimes understood as being synonymous with empathy (Lieberman, 2007) (and so sometimes is called *cognitive empathy* (Lamm et al., 2007)). From this point of view adopting another person's perspective is equivalent to a person trying to place themselves in the same psychological space as the other person, including emotions, constraints etc. and this is called the simulation theory of ToM (Goldman, 2005) (contrast with theory-theory ToM). This article proposes a simplified form of the simulation theory of ToM: an artificial mind can potentially contain a model of another agent's psychological perspective (either artificial or human), and can use this perspective to improve their decision-making in social contexts.

2.1 A 'Game Theory of Mind'

Arguably one of the most significant insights to come from economics is to ask the central question: How do people make decisions in the context of other people's decisions? This strikes close to issues central to our ToM, if one person understands that another person's actions will change the reward they will receive, then understanding the way in which that other person chooses their actions would be an invaluable tool. From this point of view, without comprehending another person's inner cognitive workings when the value of a reward depends upon the other's decisions a vast world of cooperative and competitive advantage is lost to us. Such reasoning requires individuals to account for how other's view each of the likely decision's everyone else will make, and in doing so they collectively change the decision-making patterns of the collective.

In a similar vein, decision-making has been modelled across large populations using stochastic differential equations in order to explain their choices in economic games. In a notable study, at an economic conference on game theory, the participants were asked to play a game called the *Traveler's Dilemma* for real monetary rewards (Goeree and Holt, 1999; Anderson et al., 2004). While the participants did not play the strict equilibrium strategies predicted by classical economics, they did make decisions that collectively were in agreement with a form of bounded rationality equilibrium described by statistical evolutionary equations and whose stationary states are again reminiscent of some probabilistic models used in Artificial Intelligence (AI) and theoretical psychology.

In the final study considered here, Yoshida et al. (Yoshida et al., 2008) recently proposed what they called a *Game Theory of Mind* that uses bounded recursion (of the sort: I'm thinking of you thinking of me thinking of you etc. truncated at a certain level) and value functions attributable to different players in order to model depth of strategic reasoning.

2.2 A Stochastic Model of Decision-making

Conventional game theory begins with a number of players and the rewards they receive for their joint actions, in the simplest case there are two players (here called A and B), each of which has two choices and the payoff matrices are given by:

$$\alpha = \begin{bmatrix} \alpha_1^1 & \alpha_2^1 \\ \alpha_1^2 & \alpha_2^2 \end{bmatrix}, \qquad (1)$$

$$\beta = \begin{bmatrix} \beta_1^1 & \beta_2^1 \\ \beta_1^2 & \beta_2^2 \end{bmatrix}. \qquad (2)$$

The expected utility to each player is given in terms of the joint probability $p(\alpha^i)q(\beta^j) = p(A^i)q(B^j) = p_i q_j$. Here, A denotes the player, A^i denotes a decision variable that can take two different value, either α^1 or α^2 denoting the first or second row in equation 1, and the A^i are sometimes called cumulative decision variables (see below for where this comes from) and $p(A^1) = A^1/(A^1 + A^2)$ etc. (equiv. for B), so:

$$E_a(u) = \sum_{i,j} p_i q_j \alpha_j^i, \quad E_b(u) = \sum_{i,j} p_i q_j \beta_i^j \qquad (3)$$

A further set of equations are needed as well, called the conditional expected utilities (Wolpert et al., 2012):

$$E_a(u|\alpha^i) = \sum_j q_j \alpha_j^i, \quad E_b(u|\beta^j) = \sum_i p_i \beta_i^j \qquad (4)$$

in which the expected utility to α is conditional upon α fixing their choice to α^i (similarly for β). Using these expressions we can represent how a single player (A in what follows) models the theirs and their opponents incremental changes in the decision variables during time interval dt in terms of a deterministic *drift* term and a stochastic *diffusion* term (Bogacz et al., 2006):

$$dA^i = (\alpha_1^i \widetilde{B}^1 + \alpha_2^i \widetilde{B}^2)dt + \sigma_A^i dW_A \qquad (5)$$

$$d\widetilde{B}^j = (\widetilde{\beta}_1^i A^1 + \widetilde{\beta}_2^i A^2)dt + \sigma_B^j dW_B \qquad (6)$$

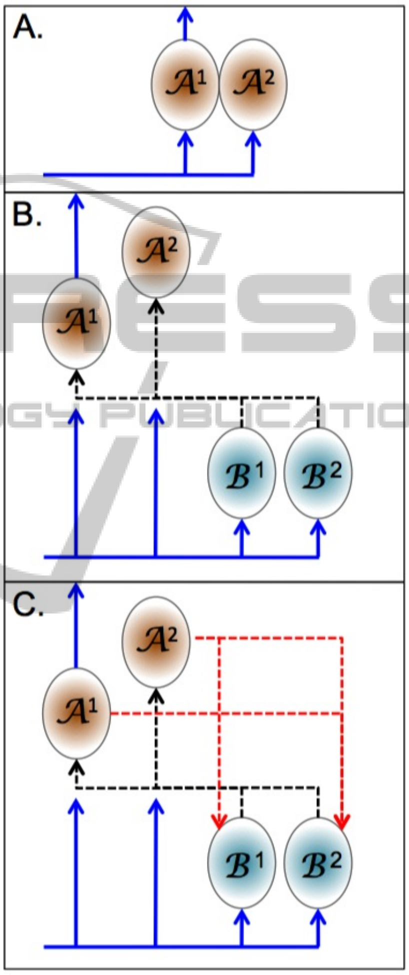

Figure 1: The neuro-cognitive development of a ToM by player A. In each panel A^1 signals before A^2 and A^1 then laterally inhibits A^2 from signalling (inhibition connections not shown). A: Two cell assemblies encode stochastic decision variables A^1 and A^2 that take early signals from other regions of the brain. B: Two new cell assemblies are formed that encode the decision variables B^1 and B^2 of another player, these decision variables influence the A^i decision variables via black dashed connections. C: The A^i are now reciprocally connected to the B^j through red dashed connections such that dynamic changes in the activity of the A^i cell assemblies during the decision making process are reflected in the activity of the B^j cell assemblies.

Note the tilde in these equations indicating that it is the A player's estimate of the B player's utility and decision variable B, in game theory the utility is common knowledge between players, but in representing the other player's decision process each player has to estimate the utility the other player will gain from the interaction. Equations 5 and 6 represent a model of how A has neurally encoded the constraints, their estimates of B's incentives as well as their decision making process (it is the same as A's). In this representation of stochastic neural signal accumulation, the dA^i are only implicitly connected to each other through their connection with the B^j (no A^i term appears on the right hand side of equation 5), however each A^i interacts with both B^j terms. This is how the neural network of A looks once it has fully developed, but the intermediate steps to this final form can be described as:

$$\begin{aligned} A.\quad dA^i &= \mu_i dt + \sigma_A^i dW_A & (7)\\ B.\quad dA^i &= (\alpha_1^i \widetilde{B}^1 + \alpha_2^i \widetilde{B}^2)dt + \sigma_A^i dW_A & (8)\\ d\widetilde{B}^j &= \widetilde{\mu}_j dt + \sigma_B^j dW_B & (9)\end{aligned}$$

where the letter labels on the left reflect the panel labels of Figure 1 and panel C. is modelled by equations 5 and 6. The μ_i term of equation 7 represents the weight the player attributes to each of their two options, because the player is not accounting for the other player's strategies at this point, a plausible (but by no means the only) strategy is to take the average of the two payoff's available for each choice for example: $\mu_i = (\alpha_1^i + \alpha_2^i)/2$ (similarly for B's μ_j). In terms of strategic thinking (see as an example (Coricelli and Nagel, 2009)), option A. is *level 0* thinking, no account is made of the other player's choices, the choice of $\mu_i = (\alpha_1^i + \alpha_2^i)/2$ implies A assumes the other player equally weights their choices, but there are obviously many other alternative formulations of μ_i. Option B. is *level 1* thinking, some account is made of B's strategy, but no attempt is made by A to adjust their interpretation of B's strategy by accounting for how B might be strategically weighting their choices based upon A's likely strategy. Finally, equations 5 and 6 represent A accounting for their estimation of the weighted strategies of B where B's weighted strategies accounts for A's weighted strategies (*level 2* thinking).

It is not easy to see that there is necessarily a solution to these equations such that an equilibrium in the dynamics might be achieved. However, it has recently previously been shown that providing the drift terms are linear in the decision variables (the terms that precede dt in equations 5- 6) there is a guaranteed set of equilibrium probabilities given by:

$$\begin{aligned} p(A^i) &\propto e^{\gamma_a E_a(u|\alpha^i)} & (10)\\ p(B^j) &\propto e^{\gamma_b E_b(u|\beta^j)} & (11)\end{aligned}$$

where the γ_a and γ_b terms are proportional to the noise terms σ in equations 5 and 6 and the remainder of the terms in the exponents are simply the conditional expected utilities of equation 4. The interpretation of this form is that if A holds strategy A^i fixed then the probability of choosing strategy A^i is proportional to the exponentiation of the utility condition on A^i being fixed based upon A's estimate of the distribution over strategies B will choose.

In order to see how equations 10 and 11 can be thought of as *social perspective taking*, simplify these probabilities to only one variable for each player: $Q_a = 1 - 2p(A^1) \in [-1,1]$ and $Q_b = 1 - 2q(B^1) \in [-1,1]$. Now rewrite equations 10 and 11 in an explicit form in terms of an equilibrium involving only Q_a for A:

$$\begin{aligned} Q_a &= 1 - \frac{2e^{\gamma_a E_a(u|\alpha^1)}}{e^{\gamma_a E_a(u|\alpha^1)} + e^{\gamma_a E_a(u|\alpha^2)}} & (12)\\ &= \tanh\left(\frac{\gamma_a}{2}(E_a(u|\alpha^2) - E_a(u|\alpha^1))\right) & (13)\\ &= \tanh\left(\frac{\gamma_a}{2}(z_0^a + z_1^a Q_b)\right) & (14)\\ &= \tanh\left(\frac{\gamma_a}{2}(z_0^a + \right.\\ &\qquad \left. z_1^a(\tanh(\frac{\gamma_b}{2}(z_0^b + z_1^b Q_a))))\right) & (15)\end{aligned}$$

The z terms are reduced constants derived from payoff matrices 1 and 2. Note that Q_a can be written as a function of Q_b: $Q_a = F_a(Q_b)$ where $F_a(\cdot)$ is A's decision model with A's γ and z parameters, cf. equation 14 and likewise $Q_b = F_b(Q_a)$. However, these are implicit self-consistent equations for each player's choices: $Q_a = F_a(F_b(Q_a))$ (cf. equation 15) and $Q_b = F_b(F_a(Q_b))$.

2.3 Social Networks, Decisions and ToM

Between players A and B a minimal 2-person social network has been described, one of mutual and self consistent comprehension between the two players (when in equilibrium) based upon an accurate mental model each has of the other. This small social network can be expanded to multiple agents, to do so label each new agent C, D, E etc. and like A and B these new comers only have two options in a game theory-like setting to choose from so that Q_c, Q_d, Q_e etc. can be defined similarly to $Q_a = F_a(Q_b)$. Taking the perspective of how A needs to internally model their social network in order to make decisions that are 'in equilibrium', A's model will depend upon the local topology of their social network. For example if A is connected to four other people and these people do not know each other or anyone else other than A, then A can represent them as independent: $Q_a = F_a(F_b(Q_a), F_c(Q_a), F_d(Q_a), F_e(Q_a))$, i.e.

$$\begin{aligned} Q_a \quad = \quad & \tanh\Big[\frac{\gamma_a}{2}(z_0^a + \\ & z_1^a(\tanh(\frac{\gamma_b}{2}(z_0^b + z_1^b Q_a))) + \\ & z_2^a(\tanh(\frac{\gamma_c}{2}(z_0^c + z_1^c Q_a))) + \\ & z_3^a(\tanh(\frac{\gamma_d}{2}(z_0^d + z_1^d Q_a))) + \\ & z_4^a(\tanh(\frac{\gamma_e}{2}(z_0^e + z_1^e Q_a))))\Big] \qquad (16) \end{aligned}$$

where the z and γ notation has been extended to these new players, see Figure 2, B. Alternatively, given three agents that all know each other and interact with each other, A's model of their local social network would be $Q_a = F_a(F_b(F_c(Q_a)))$, i.e.

$$\begin{aligned} Q_a \quad = \quad & \tanh\Big[\frac{\gamma_a}{2}(z_0^a + \\ & z_1^a(\tanh(\frac{\gamma_b}{2}(z_0^b + \\ & z_1^b(\tanh(\frac{\gamma_c}{2}(z_0^c + \\ & z_2^c(\tanh(\frac{\gamma_d}{2}(z_0^d + z_1^d Q_a)))\Big] \qquad (17) \end{aligned}$$

see Figure 2, C. In this example it would not be a complete representation of the local network topology for A to simply have accounted for B and C as though these two people were independent of each other, in such a model: $Q_a = F_a(F_b(Q_a), F_c(Q_a))$, but this does not accurately reflect how A needs to consider the way in which B and C adjust their decisions based upon the connection they have with each other, as this connection indirectly influences how A needs to consider the choices they make. More generally, in the case of B. of Figure 2, while A is influenced by B (and C, D and E), B might themselves be connected to other players in their local network and these players are only indirectly related to A through the influence they have on B's decisions.

3 CONCLUSIONS

How do agents, artificial or biological, coordinate their actions in such a way that their collective decision-making is better than their individual decision-making? People are both good and bad at such aggregation: science is based on individuals cooperating and competing in order to produce a body of knowledge far greater than any one individual could achieve, but the coordinated behaviour of traders in financial markets can lead to billions of dollars being lost in a market crash. More importantly for artificial agents, what are the theoretical foundations of cooperation and competition that can lead to simple agents making better decisions collectively than could be achieved through every individual acting independently of one another? So how do we engineer true collective intelligence in societies of artificial agents?

One approach is outlined in this work: individual agents make their decisions based on what they can each individually and objectively know (or can find out) about the environment as well as how they believe other agents will make their choices and how these choices subsequently impact on the quality of their own decisions. In people, such cooperative division of labour leads to specialisation and expertise amongst a population that is able to solve problems that no individual could solve alone, replicating this in a collection of artificial intelligences will open up the possibilities of a symbiosis of bio-silica communities, where artificial agents are dynamically responsive to the internal mental states of people, enhancing the quality of our decision-making.

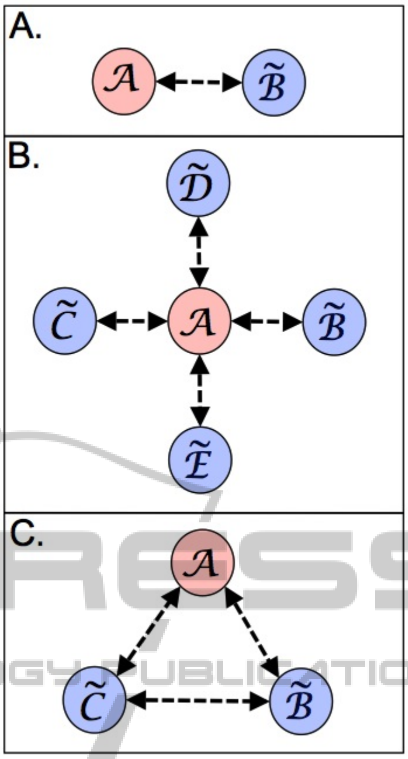

Figure 2: The internal representations A has of the other agents in different network topologies. Each topology has a distinctive representation when A's choices are in equilibrium: A: $Q_a = F_a(F_b(Q_a))$, B: $Q_a = F_a(F_b(Q_a), F_c(Q_a), F_d(Q_a), F_e(Q_a))$, C: $Q_a = F_a(F_b(F_c(Q_a)))$.

REFERENCES

Anderson, S. P., Goeree, J. K., and Holt, C. A. (2004). Noisy directional learning and the logit equilibrium. *The Scandinavian Journal of Economics*, 106(3):581–602.

Baron-Cohen, S. E., Tager-Flusberg, H. E., and Cohen, D. J. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience* . Oxford University Press.

Bhatt, M. and Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: fmri evidence. *Games and Economic Behavior*, 52(2):424–459.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700.

Coricelli, G. and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23):9163–9168.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.

Dunbar, R. I. and Shultz, S. (2007). Evolution in the social brain. *science*, 317(5843):1344–1347.

Goeree, J. K. and Holt, C. A. (1999). Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567.

Goldman, A. I. (2005). Imitation, mind reading, and simulation. *Perspectives on Imitation: Imitation, human development, and culture*, 2:79.

Harré, M. (2013). The neural circuitry of expertise: Perceptual learning and social cognition. *Frontiers in Human Neuroscience*, 7:852.

Kudo, H. and Dunbar, R. (2001). Neocortex size and social network size in primates. *Animal Behaviour*, 62(4):711–722.

Lamm, C., Batson, C. D., and Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of cognitive neuroscience*, 19(1):42–58.

Lehmann, J. and Dunbar, R. (2009). Network cohesion, group size and neocortex size in female-bonded old world primates. *Proceedings of the Royal Society B: Biological Sciences*, 276(1677):4417–4422.

Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, 58:259–289.

Powell, J., Lewis, P. A., Roberts, N., García-Fiñana, M., and Dunbar, R. (2012). Orbital prefrontal cortex volume predicts social network size: An imaging study of individual differences in humans. *Proceedings of the Royal Society B: Biological Sciences*, 279(1736):2157–2162.

Shultz, S. and Dunbar, R. (2007). The evolution of the social brain: Anthropoid primates contrast with other vertebrates. *Proceedings of the Royal Society B: Biological Sciences*, 274(1624):2429–2436.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Wolpert, D. H., Harré, M., Olbrich, E., Bertschinger, N., and Jost, J. (2012). Hysteresis effects of changing the parameters of noncooperative games. *Physical Review E*, 85(3):036102.

Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS computational biology*, 4(12):e1000254.

Zacks, J. M. and Michelon, P. (2005). Transformations of visuospatial images. *Behavioral and Cognitive Neuroscience Reviews*, 4(2):96–118.