games

MDPI

Opinion

What Can Game Theory Tell Us about an AI 'Theory of Mind'?

Michael S. Harré

Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Sydney 2006, Australia; michael.harre@sydney.edu.au

Abstract: Game theory includes a rich source of methods for analysing strategic interactions where there are a small number of agents, each having only a few choices. In more complex settings though, where there are many choices over indefinite time horizons involving large social groups, these methods are unlikely to fully capture the causes of agent behaviour. If agents are able to simplify the task of understanding what others might do by modelling the constraints of others, particularly unobservable *cognitive* constraints, then the possible *behavioural* outcomes can be similarly restricted, thereby reducing the complexity of a social interaction. Having a cognitive representation of the unobserved causal states of others is an aspect of a 'Theory of Mind' and it plays a central role in the psychology of social interactions. In this article I examine a selection of results on the theory of mind and connect these with the 'game theory of mind' to draw conclusions regarding the complexity of one-on-one and large-scale social coordination. To make this explicit, I will illustrate the relationship between the two psychological terms 'introspection' and 'theory of mind' and the economic analysis of game theory, while retaining as much as possible of the richness of the psychological concepts. It will be shown that game theory plays an important role in modelling interpersonal relationships for both biological and artificial agents, but it is not yet the whole story, and some psychological refinements to game theory are discussed.

Keywords: game theory; psychology; theory of mind; introspection; artificial intelligence; social networks; strategic cooperation; strategic threat; adversarial modelling

Citation: Harré, M.S. What Can Game Theory Tell Us about an AI 'Theory of Mind'? *Games* **2022**, *13*, 46. https://doi.org/10.3390/g13030046

Academic Editor: Ulrich Berger

Received: 30 April 2022
Accepted: 16 June 2022
Published: 20 June 2022

1. Introduction

The coordination between groups of artificial agents and humans will become one of the most challenging tasks for improving AI's ability to support human endeavours [1,2]. In order to do this, we first need to understand how intelligent agents make choices in the context of other intelligent agents, a complex task for which humans have specific and highly developed cognitive processes that are detectable even in early childhood [3]. For example, in-group social dynamics are caused by people collectively orientating themselves towards 'like minded' others in social group formation (otherwise known as intra-group homophily) [4]. This reduces the cognitive load of having to try and understand everyone else in the group by assuming that everyone, to some reasonable approximation, is like everyone else. It also plays an important role in threat assessment: knowing the causal constraints of other agents, or collections of agents, helps to narrow the range of possible threats they pose, such antagonistic inter-group heterogeneity with intra-group homogeneity is often the source of the echo-chamber effect in social media, as shown by Colleoni et al. [5]. Notably, this study also showed that the individual network structures of different homophylic groups can be very different, suggesting that there is something in the specifics of the homophilous connections that influences social network topology (cf. Section 3). With both of these elements there is a need to better understand the mechanisms of both cooperation for achieving joint goals and competition for analysing strategic threats.

In a certain sense, two-agent games contain the basic foundations for experimental and theoretical studies into social interactions as well as forming the simplest social network containing two nodes (players or agents), with a bi-directional link between the nodes

indicating a relationship exists from one node to the other. Despite its simplicity, this arrangement has provided a wealth of insights into the nature of collective behaviour of not just humans but also, for example, the behaviour of other species and how genes interact with one another. However, recent new developments have motivated novel analyses in game theory, three of which are discussed in this opinion piece: theory of mind (ToM), social network theory, and how game theory can connect these to develop new ideas in artificial intelligence (AI). In order to situate game theory in a psychological framework, a brief discussion of the meaning and significance of individual cognition and interpersonal cognition is introduced next before moving to the main topic of this article: the psychology of a 'Game Theory of Mind'.

1.1. Individual Cognition

On the left in Figure 1 is Barrett's ([6], Figure 1) analysis of recent research in cross-cultural cognitive sciences, illustrating the relationships between fields of study in the cognitive sciences. Individual cognition has also been studied recently by Peterson et al. [7] in the context of gambling and uncertainty. Some of the models they analysed are on the right of Figure 1 (simplified from Figure 1c in [7]). Peterson et al showed that well-known computational theories, such as expected value theory, prospect theory, and many others, are empirically sound using very large databases of people's decisions. These results tell us not only what is necessary to accurately represent individual cognitive processes, but also how sophisticated a biological agent's cognitive model of others needs to be in order to, at least partially, understand their individual cognitive processes.

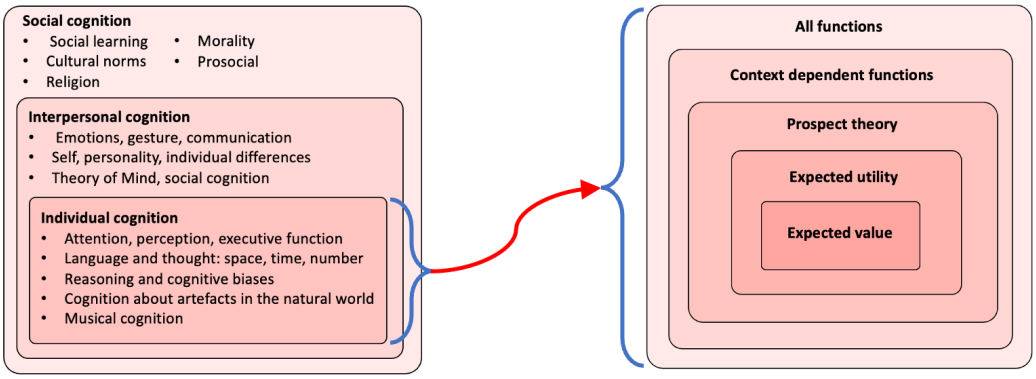

Figure 1. (**Left**) a simplified form of the psychological framework of Barrett [6] illustrating the multi-modal nature of cognition, the individual, the interpersonal, and the social. (**Right**) A simplified form of Peterson et al.'s [7] layering of cognitive models for AI. Note that this is, in part, an AI representation of well known socio-economic decision models.

1.2. Social Constraints

At the highest level of Barrett's analysis is social cognition, which includes the elements of moral, cultural, and religious norms. In another recent result using large data analysis, Awad et al. [8] developed a questionnaire for what they called the Moral Machine in which they gathered 40 million decisions from participants from 233 countries. What they found is significant cross-cultural ethical variation with three large clusters of countries and that these differences correlate with institutions and cultural traits.

From a certain perspective, it may not be surprising that people with different backgrounds are different from one another, but this would miss two insights: (1) it is possible to quantify these differences; (2) that these differences have a measurable impact on choices in ambiguous scenarios. We can think of these large-scale patterns of behaviour as providing constraints on the decisions people make at the individual level when there is no strict logic, that is to say that knowing where someone comes from reduces the ambiguity another agent would have regarding their possible actions.

1.3. Theory of Mind and Introspection

A person is said to have a ToM if they are able to impute mental states, such as desires, intentions, and beliefs, to another person—see, for example, Frith and Frith [9] for a short summary. The emergence of a ToM is a key milestone in the interpersonal development of children [10], and the failure for it to develop is associated with autism, which, as a diagnosis, is in part characterised by a difficulty in understanding social contexts [11].

Introspection, on the other hand, is the facility to understand one's own thoughts, for example to have a representation of our own thought processes. Recent work in this area has combined introspection with brain-mapping [12], and for an early historical review see Boring [13]. Introspection is also closely related to ToM, they are both central to early childhood development and there is an ongoing debate regarding how introspection and ToM interact with each other during early development; see, for example, Gonzales et al. [14] for a recent discussion. Figure 2 illustrates the relationship between introspection and ToM, and it will be shown that when they interact within a single agent the results have surprising complexities.

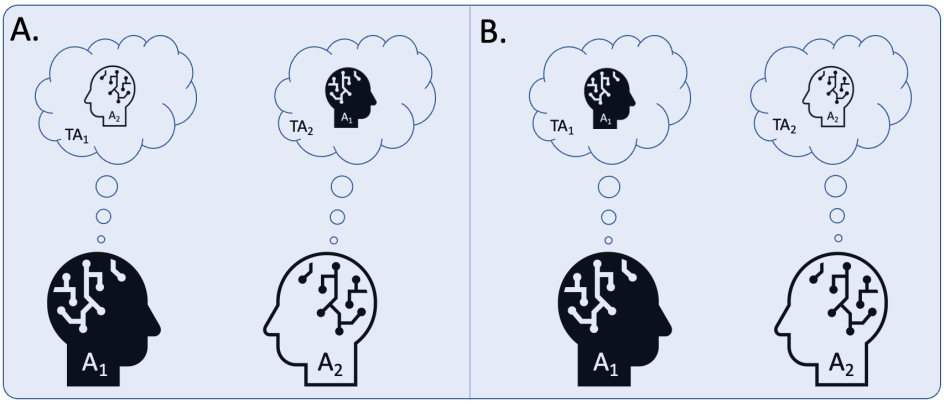

Figure 2. (**A**) both agents A_1 and A_2 have a representation of the other agent's cognition (ToM). (**B**) both agents have a representation of their own cognition (introspection). TA_i is a cognitive space (see, for example, [15,16]) in which the 'thoughts of agent i' occur.

What follows in this article is an indication of the complexity of the relationship between certain psychological phenomena and their possible implementation in a working, socially adept AI where game theory underpins the theoretical aspects. This goes some way to addressing the concerns Shevlin and Halina [17] have in which 'rich psychological terms' should not misused. In what follows are two main sections. The first expands on introspection and ToM in a way that makes some of their key features operational in an AI, while illustrating the relationship with game theory. The second then discusses how this is related to building and maintaining complex social networks. In the final section I discuss areas for future work.

2. The 'Game Theory of Mind': Neuroscience and Economics in Strategic Interactions

The game theory of mind (GToM) model was introduced in a 2008 paper by Yoshida et al. [18] to describe how people represent the causal cognitive states of others in order to optimise interpersonal interactions. The premise is based on cooperating or defecting in a hunting game: an agent can hunt collectively with other agents, i.e., be socially cooperative, and earn a large reward by killing a large animal, or they can hunt as an individual, i.e., defect from the group, and earn a smaller reward. The goal is to develop a formal description, based on empirical evidence from human fMRI data, of "...a model of 'theory of mind' using 'recursive sophistication' in which my model of your goals includes a model of your model of my goals, and so on ad infinitum". The people in the experiment played against a computer with two different strategies: a lower order competitive strategy and a higher order cooperative strategy, and the premise is that game theory can elucidate

how agents build internal representations of strategic levels of social interactions from observations of their opponent's behaviour. Experiments such as this have been extended beyond behavioural models to directly studying the behavioural and neural recordings in monkeys [19–24] and humans [25–30]. These and many other studies have significantly contributed to our understanding of how we attribute value to our decisions, where this is processed in the brain, and how our allocation of value to the decisions of others interacts with the value we allocate to our own decisions.

In order to accommodate these facets, research such as that of Griessinger and Coricelli [31] and Yoshida et al. [32] have posited models that allow for varying degrees (depth) of strategic reasoning. For example in Griessinger and Coricelli [31] they modelled reinforcement learning (level 0 strategic reasoning), fictitious play learning (level 1 strategic reasoning), and influence learning (level 2+ strategic reasoning). Much of this work has used the notion of strategic depth, the extent to which one agent reasons about the other agent reasoning about them, etc., as illustrated in Figure 3 (bottom, A.), which does indeed have elements of ToM. There is very little discussion of the relationship between ToM, introspection, and strategy though, arguably because it is not clear where introspection sits in an analysis of games despite its distinct psychological role independent of ToM. See, for example, Goeree and Holt [33], where introspection corresponds to ToM as Yoshida et al. [18] define it. Figure 3 shows where this distinction becomes important, not only does an agent A_1 need to be aware of A_2's reasoning about A_1 (top row, A.), but A_1's representation of the strategic interaction needs to include a representation of their own reasoning about the situation (top row, B.). With a model of their own cognitive abilities, A_1 can understand, for example, its limited understanding of A_2. The second row of Figure 3 shows the next step in deepening this strategic thinking.

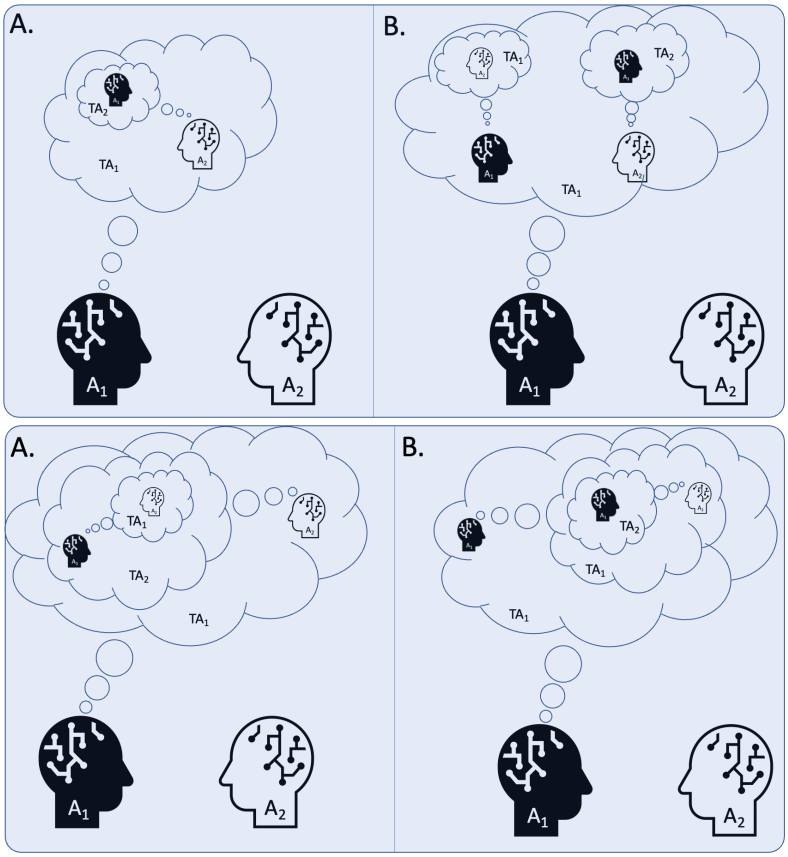

Figure 3. Top row, (**A**) two levels of strategic thinking: A_1 thinking about A_2, thinking about A_1. Top row, (**B**) A_1 has two distinct representations, one in which A_1 is introspecting on their model of how A_1 reasons about A_2 and the other is a A_1's ToM of how A_2 is thinking about A_1. Bottom row, (**A**) the next strategic level based on ToM. Bottom row, (**B**) The next strategic level based on introspection.

To make the case more explicit, these concepts can be re-framed in terms related to game theory. In the work of Yoshida et al. [18,32], they modelled levels of cognitive reasoning but they also connect it, in a suitable limit, to the quantal response equilibrium (QRE) of McKelvey and Palfrey [34]. It is important to note that the QRE is explicitly an equilibrium model, and it contains within its limits the conventional notion of a Nash equilibrium, but we can also use it to illustrate how ToM relates to a specific GToM for two-player games, as follows. Given two players a and b each with a binary strategy set $S_i = \{s_i^1, s_i^2\}$, $i \in \{a, b\}$, the QRE can be represented as a self-consistent mapping ($-i$ denotes i's opponent). We supplement the usual treatment of this model by assuming that i only possesses an approximate representation of $-i$'s decision model and subsequent behaviour, and we denote this internal representation with a prime on those terms that are internal to i rather than 'true' representations of the other agent:

$$\begin{aligned} \bar{s}_i &= p(s_i^1|\bar{s}'_{-i}) \propto \exp(\beta_i E(u_i|\bar{s}'_{-i})) && (1)\\ \bar{s}_i &= \psi_i(\bar{s}'_{-i}|c_i) && (2)\\ \bar{s}_i &= \psi_i(\psi'_{-i}(\bar{s}_i|c_{-i})|c_i) && (3) \end{aligned}$$

In the case where $\bar{s}'_{-i} = \bar{s}_{-i}$ and $\psi'_{-i}(\bar{s}_i|c_{-i}) = \psi_{-i}(\bar{s}_i|c_{-i})$ Equation (3) defines the original QRE model [34]. In Equation (1) $\bar{s}_i \in [0, 1]$ is the probability of player i playing strategy 1, $p(s_i^j|\bar{s}'_{-i})$ is the probability of player i choosing strategy j subject to agent i's approximation of $-i$'s choosing strategy 1, $E(u_i^j|\bar{s}'_{-i})$ is the expected utility to i for playing j subject to $-i$'s probability of play, and β_i is the rationality, or 'persona' [35], of agent i. There are also payoff matrix parameters (encoded in the expected utility function) and these, along with the β_i, are collected in the co-vector of i's parameters denoted c_i. In Equation (3) we simply express the notion that i's probability is a function of $-i$'s probability. Also note that this logit form of noisy decision-making is the game-theoretical analogue of the expected utility function Peterson et al. [7] used for individual cognition, discussed in the Introduction above.

While there is nothing novel about this representation, as the QRE has been used extensively as a model of bounded rationality [36–40], there are nuances that are worth making explicit. First note that Equation (2) has only a behavioural interpretation of the other player. Viewing $\bar{s}_{-i}$ as the behaviour of an inanimate object, the agent learns about S'_{-i} but it is limited in that $\bar{s}'_{-i}$ does not have any sophisticated internal causal structure. In this interpretation, the agent only learns through observation the probabilities over possible outcomes and then, based on the rewards they receive, constructs a probability for their own choices: $\bar{s}_i$. However, once an agent learns that $\bar{s}_{-i}$ might be similar to themselves and has some form of internal structure of its own then Equation (3) illustrates how an agent can use this representation to modulate their own decision-making based on the other agent's constraints c_{-i}.

Now suppose that i has an internal representation of the strategic interaction that i is involved in, i.e., Equations (1)–(3), but the internal representation is only an approximate representation, so we denote each term with a prime: $\bar{s}'_i$, $\psi'_i(\cdot)$, and $\psi'_{-i}(\cdot)$, $\longleftarrow$ denotes the direction of causation i has learned in constructing this representation, and $\longleftrightarrow$ denotes a correspondence between two cognitive representations. The internal representation i has of the strategic interaction can be written in the following ways:

$$\Gamma^i : \begin{cases} \bar{s}'_i & \longleftarrow \psi'_i(\psi'_{-i}(\bar{s}'_i|c_{-i})|c_i) \\ {\psi'_i}^{-1}(\bar{s}'_i|c_i) & \longleftrightarrow \psi'_{-i}(\bar{s}'_i|c_{-i}) \end{cases} \quad (4)$$

and Γ^i is A_i's encoding, in either an artificial or a biological neural network. In the first line of Equation (4) is the standard representation of the QRE. The second line is the correspondence between introspection on the left, given by the inverse of A_i's representation of themselves: ${\psi_i}^{-1}(\cdot)$, and A_i's ToM representation of A_{-i}: $\psi_{-i}(\cdot)$. Schematically, line one of Equation (4) is illustrated in Figure 4A. (introspection, the left-hand representation in

A_1's thought bubble where A_1 is thinking of A_2's strategy) and line two is illustrated in Figure 4B. (the correspondence between ToM and introspection). Line two is an algebraic manipulation that makes explicit that A_i has two distinct models, one of itself and the other of A_i. In practice ${\psi_{i'}}^{-1}(\cdot)$ may not be the inverse of the QRE as it can be any encoding that A_i has of its own individual or interpersonal cognition (c.f. Figure 1), but this representation is in an equilibrium correspondence with A_i's encoding of A_{-i} in models such as the QRE.

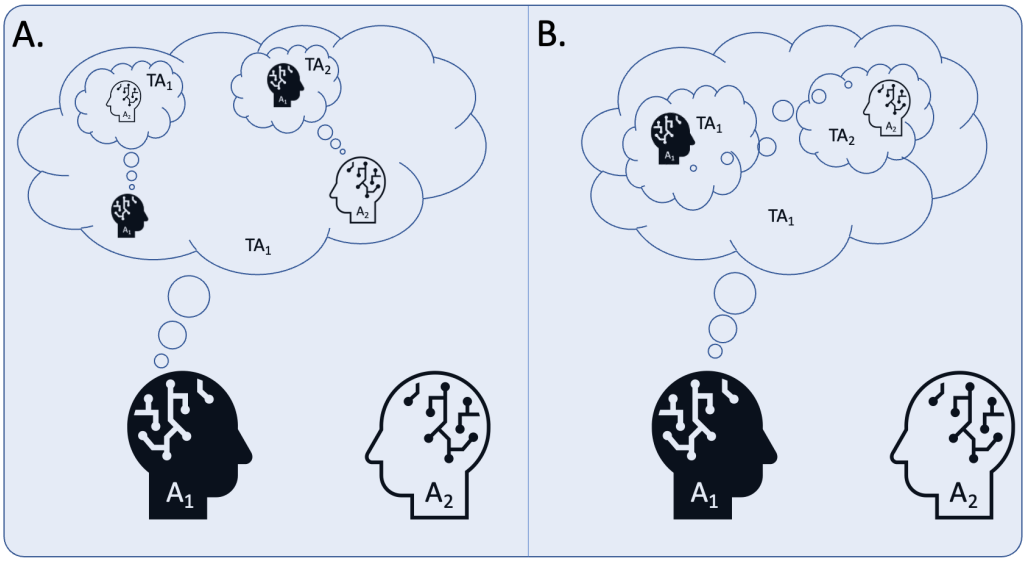

Figure 4. (**A**) A_1 has a representation of themselves (via introspection), that includes A_2 as well as a representation of A_2 that includes A_2's representation of A_1 (via ToM). (**B**) The QRE model requires that the two distinct representations in (**A**) are in correspondence with one another, thereby making A_1's strategy self-consistent with A_1's representation of A_2.

Aspects of this formal approach have been developed into experiments in which the results support a strategic role for our ToM. In a 2004 study, Rilling et al. [41] used the Prisoner's Dilemma (PD) and the Ultimatum Game (UG) in an fMRI study of the parts of the brain that were active in two different types of strategic interaction and if these regions corresponded to the known neural network that is activated during other ToM tasks. In both games, regions of the brain associated with a ToM were activated when humans were told they were playing both another human and a computer, but in the case of humans the activation was much stronger, suggesting that we differentiate our thinking of our opponents depending on what we know of their possible internal states, where the authors concluded: 'The fact that computer partners are able to activate this network, albeit to a lesser extent than human partners, suggests that either this neural system can also be activated by reasoning about the unobservable states of nonhuman systems, or that participants imbue their computer partners with human attributes'. There were further differences between the neural activity of the test subjects that separated the UG from the PD, and this no doubt also points to subtle cognitive differences in how humans respond to strategic interactions based on the different hidden states of either their computer or human opponent. However, the final point I would like to make is that the QRE model, specifically its 'persona' version [35] in which the covector of parameters encodes the persona of the players, goes a significant way to explaining the differences between experimental results and the Nash equilibrium of the UG [42]. The interpretation of this result is that, in order to avoid subsequent punishment for an unfair split in this game, the agent who splits the value needs to take into account how likely the other agent is to punish them for an unfair split, and this pre-emptive accounting of future punishments is a type of understanding of the internal constraints the receiver has that the splitter needs to account for. These results have subsequently been borne out in subsequent studies as well [43–45].

3. The Importance of a Theory of Mind in Human-to-Human Interactions

In the field of anthropology there is a well-studied relationship between the surface area of the human brain and the size of our largest stable social grouping, approximately 150 individuals. This number, called Dunbar's number [46,47], is also known to have a number of highly stratified layers within this grouping that cluster around mean values of ≃1.5, 5, 15, 35, and 150 people (note that there are multiple views on the exact numbers) [47–49]. These numerical values are related to the variety of cognitive adaptations people use to maintain long-term stable relationships, such as pair bonding, intimate friendships, casual acquaintances, and more distant associations. In particular, these numbers are a consequence of the topology of the social networks humans participate in, and in this case these are of the Erdös–Rényi variety of networks [49]. This topology is driven by people forming discrete links between one another that subsequently constrains the network topology. Notably, humans have the largest social network of any other comparable species [46], and at the same time humans have the largest ratio of neocortex size to total brain volume [50], and it is from this relationship that the Social Brain Hypothesis was developed to explain why the human brain has been so successful in manipulating the natural environment: it is due to human's ability to develop and maintain complex social interactions that in turn leads to sophisticated solutions to tasks in the natural environment.

The inner layers of our social networks (called the support clique and the sympathy group) have been shown to be mediated by our ability to form social connections based on our capacity to use ToM and memory in order to form connections between individuals. In an MRI study, Powell et al. [51] showed that the volume of the orbital pre-frontal cortex positively correlates with the size of a subjects' inner social network layers. This built on the earlier work of Stller and Dunbar [52] who showed that memory helps explain the sympathy group (the social layer of approximately 12–15 individuals) and perspective taking, i.e., ToM, helps explain the support clique of approximately 4–7 group members. As a final example, in the study by Lewis et al. [53] it was shown that an individual's social network size is predicated upon the volume of neural material dedicated to specific types of mentalising.

The cognitive complexity of maintaining these networks is illustrated in Figure 5, each of the links $L_{1,i}$ that A_1 forms in the real world needs to have a cognitive representation $TL_{1,i}$ [16] and these links are not usually independent of one another. As mentioned above we also have specific strategies for managing these complexities and, to the extent that we will need socially adept AI, the next generation of AI agents will also need to have better internal representations of their human partners, how these partners see them, and how this influences collective behaviours.

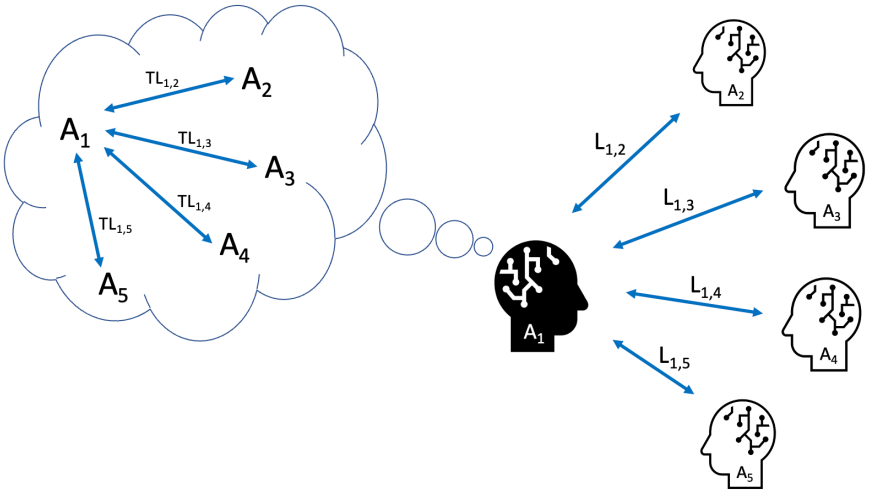

Figure 5. Real social networks and the cognitive representation needed for an agent to successfully navigate the social world.

4. Conclusions

Introspection and Theory of Mind are psychologically distinct, complex notions that are tightly inter-related, both developmentally and in later life. In this article I have used these two concepts to illustrate how a more precise use of psychological ideas can lead to a more nuanced understanding of strategic interactions, to look more closely 'under the hood' of cognition [54]. For example, a key point of this article is the distinction between how an agent makes a decision: $\psi_i(\cdot)$ versus how this is cognitively encoded by an agent: $\psi'_i(\cdot)$. The first is a 'true' representation of the agent's decision-making process, the second is an approximation of themselves an agent has learned. In contrast, many 'game theory of mind' approaches take the strategic levels beginning with a true representation of A_i thinking of A_{-i} who is thinking of A_i's strategy i.e., $\psi_i(\psi'_{-i}(s_i))$. This takes the first level, $\psi_i(\cdot)$, to be how A_i makes a decision, and this then contains a representation of agent A_{-i}'s decision-making process: $\psi'_{-i}(\cdot)$. In this article, an agent can reason about their own cognition via introspection or about another agent's cognition via ToM. The former is $\psi_i(\psi'_i(\cdot))$ and the latter is $\psi_i(\psi'_{-i}(\cdot))$; this is the distinction made in Figure 2. The consequence is an awareness of the relationship between 'self' and 'other', the representations being manipulated by the agent, and that these representations are consistent with respect to one another in an equilibrium models such as the QRE. This allows an agent to have greater flexibility when thinking of themselves in relation to others.

A further consequence is that if the equilibrium assumption is relaxed then encodings do not need to be related to one another (c.f. Figure 4A), which can be represented as: $TA_i[\psi'_i(\cdot), \psi'_{-i}(\cdot)]$ where TA_i is the cognitive space of A_i, as described in the caption to Figure 2. These considerations bring to the fore three different notions of equilibrium: agents that are in equilibrium as in the conventional QRE (co-vectors are dropped to reduce notational clutter)—$\psi_i(\psi_{-i}(s_i)) = s_i$; an agent who is using ToM as they think of another agent—$\psi_i(\psi'_{-i}(s_i)) = s_i$; and an agent that has has adapted both a ToM and introspection to interpret the interaction—$TA_i[\psi'_i(\psi'_{-i}(s'_i)) = s'_i]$, and so providing a clearer understanding of the different psychological states of the agents that each implies. Each $\psi_i(\cdot)$ can also represent many different types of individual cognition (c.f. Figure 1), for example the seven described in Ert et al. [55] or the dozens listed in Peterson et al. [7]. The importance for the current article is that the models themselves are the constraints, they provide testable hypotheses for the limits of human cognition, how we represent ourselves and others, and how we can take advantage of these limits to understand and predict the behaviour of one another.

To make this specific in a realistically complex setting, consider the game of Go that the AI AlphaGo was so successful at playing [56]. The number of moves available at each stage of the game ranges from a relatively small number towards the end of the game to theoretically hundreds at the beginning of the game that branch out into an incomprehensibly large move space, so how do people play games like this? Clearly not by examining all, or even a representative sample of all, the moves available to them, and likewise their (human) opponent cannot examine a representative sample of moves. Of course, an enormous amount of experience and task-specific learning is involved, but ultimately people use a relatively small number of initial positions to investigate where to move next and then they explore a relatively narrow range of future plays from that small number of starting points, and this works because both players are subject to the same (cognitive) limitations. This is a common to all game experts and consequently there is not a great deal of variation in the number of starting positions considered between chess masters and intermediate players [57]. Not surprisingly, where they differ is in the quality of the moves they consider or the style of their play. While this is a complicated topic, the recent articles by Ericsson [58] and Gobet and Charness [59] review the key developments in the history of the psychology of expertise and how performance relates to perception, a topic I explored previously using an artificial neural network to model perceptual expertise in Go [60,61].

It is only through an interdisciplinary understanding of how the many different fields have approached the problem of coordinated social behaviour, from economics, to neuroscience, to psychology, and in AI, that we can begin to model and implement these intrinsically human-like aspects with the fidelity necessary to take advantage of the benefits AI can afford us.

Funding: This research did not receive any funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wang, D.; Churchill, E.; Maes, P.; Fan, X.; Shneiderman, B.; Shi, Y.; Wang, Q. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–6.
2. Dellermann, D.; Calma, A.; Lipusch, N.; Weber, T.; Weigel, S.; Ebel, P. The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems. *arXiv* **2021**, arXiv:2105.03354.
3. Bian, L.; Baillargeon, R. When Are Similar Individuals a Group? Early Reasoning About Similarity and In-Group Support. *Psychol. Sci.* **2022**, *33*, 752–764. [CrossRef] [PubMed]
4. Halberstam, Y.; Knight, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *J. Public Econ.* **2016**, *143*, 73–88. [CrossRef]
5. Colleoni, E.; Rozza, A.; Arvidsson, A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **2014**, *64*, 317–332. [CrossRef]
6. Barrett, H.C. Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends Cogn. Sci.* **2020**, *24*, 620–638. [CrossRef]
7. Peterson, J.C.; Bourgin, D.D.; Agrawal, M.; Reichman, D.; Griffiths, T.L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **2021**, *372*, 1209–1214. [CrossRef]
8. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.F.; Rahwan, I. The moral machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef]
9. Frith, C.; Frith, U. Theory of mind. *Curr. Biol.* **2005**, *15*, R644–R645. [CrossRef]
10. Korkmaz, B. Theory of mind and neurodevelopmental disorders of childhood. *Pediatr. Res.* **2011**, *69*, 101–108. [CrossRef]
11. Hughes, C.; Leekam, S. What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Soc. Dev.* **2004**, *13*, 590–619. [CrossRef]
12. Jack, A.I.; Roepstorff, A. Introspection and cognitive brain mapping: From stimulus–response to script–report. *Trends Cogn. Sci.* **2002**, *6*, 333–339. [CrossRef]
13. Boring, E.G. A history of introspection. *Psychol. Bull.* **1953**, *50*, 169. [CrossRef] [PubMed]
14. Gonzales, C.R.; Fabricius, W.V.; Kupfer, A.S. Introspection plays an early role in children's explicit theory of mind development. *Child Dev.* **2018**, *89*, 1545–1552. [CrossRef]
15. Newby, G.B. Cognitive space and information space. *J. Am. Soc. Inf. Sci. Technol.* **2001**, *52*, 1026–1048. [CrossRef]
16. Breckler, S.J.; Pratkanis, A.R.; McCann, C.D. The representation of self in multidimensional cognitive space. *Br. J. Soc. Psychol.* **1991**, *30*, 97–112. [CrossRef]
17. Shevlin, H.; Halina, M. Apply rich psychological terms in AI with care. *Nat. Mach. Intell.* **2019**, *1*, 165–167. [CrossRef]
18. Yoshida, W.; Dolan, R.J.; Friston, K.J. Game theory of mind. *PLoS Comput. Biol.* **2008**, *4*, e1000254. [CrossRef]
19. Barraclough, D.J.; Conroy, M.L.; Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **2004**, *7*, 404–410. [CrossRef]
20. Schultz, W. Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Curr. Opin. Neurobiol.* **2004**, *14*, 139–147. [CrossRef]
21. Lee, D. Game theory and neural basis of social decision making. *Nat. Neurosci.* **2008**, *11*, 404–409. [CrossRef]
22. Camerer, C.F. Behavioral game theory and the neural basis of strategic choice. In *Neuroeconomics*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 193–206.
23. Harré, M.S. Strategic information processing from behavioural data in iterated games. *Entropy* **2018**, *20*, 27. [CrossRef] [PubMed]
24. Ong, W.S.; Madlon-Kay, S.; Platt, M.L. Neuronal correlates of strategic cooperation in monkeys. *Nat. Neurosci.* **2021**, *24*, 116–128. [CrossRef] [PubMed]
25. Montague, P.R.; Berns, G.S.; Cohen, J.D.; McClure, S.M.; Pagnoni, G.; Dhamala, M.; Wiest, M.C.; Karpov, I.; King, R.D.; Apple, N.; et al. Hyperscanning: Simultaneous fMRI during linked social interactions. *Neuroimage* **2002**, *16*, 1159–1164. [CrossRef] [PubMed]

26. Bhatt, M.; Camerer, C.F. Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games Econ. Behav.* **2005**, *52*, 424–459. [CrossRef]
27. Fukui, H.; Murai, T.; Shinozaki, J.; Aso, T.; Fukuyama, H.; Hayashi, T.; Hanakawa, T. The neural basis of social tactics: An fMRI study. *Neuroimage* **2006**, *32*, 913–920. [CrossRef]
28. Kuss, K.; Falk, A.; Trautner, P.; Montag, C.; Weber, B.; Fliessbach, K. Neuronal correlates of social decision making are influenced by social value orientation—An fMRI study. *Front. Behav. Neurosci.* **2015**, *9*, 40. [CrossRef]
29. Chen, Y.H.; Chen, Y.C.; Kuo, W.J.; Kan, K.; Yang, C.; Yen, N.S. Strategic motives drive proposers to offer fairly in Ultimatum games: An fMRI Study. *Sci. Rep.* **2017**, *7*, 527. [CrossRef]
30. Shaw, D.J.; Czekóová, K.; Staněk, R.; Mareček, R.; Urbánek, T.; Špalek, J.; Kopečková, L.; Řezáč, J.; Brázdil, M. A dual-fMRI investigation of the iterated Ultimatum Game reveals that reciprocal behaviour is associated with neural alignment. *Sci. Rep.* **2018**, *8*, 10896. [CrossRef]
31. Griessinger, T.; Coricelli, G. The neuroeconomics of strategic interaction. *Curr. Opin. Behav. Sci.* **2015**, *3*, 73–79. [CrossRef]
32. Yoshida, W.; Seymour, B.; Friston, K.J.; Dolan, R.J. Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* **2010**, *30*, 10744–10751. [CrossRef]
33. Goeree, J.K.; Holt, C.A. A model of noisy introspection. *Games Econ. Behav.* **2004**, *46*, 365–382. [CrossRef]
34. McKelvey, R.D.; Palfrey, T.R. Quantal response equilibria for normal form games. *Games Econ. Behav.* **1995**, *10*, 6–38. [CrossRef]
35. Wolpert, D.; Jamison, J.; Newth, D.; Harré, M. Strategic choice of preferences: The persona model. *J. Theor. Econ.* **2011**, *11*, 1–37. [CrossRef]
36. Wolpert, D.H.; Harré, M.; Olbrich, E.; Bertschinger, N.; Jost, J. Hysteresis effects of changing the parameters of noncooperative games. *Phys. Rev. E* **2012**, *85*, 036102. [CrossRef] [PubMed]
37. Harré, M.S.; Atkinson, S.R.; Hossain, L. Simple nonlinear systems and navigating catastrophes. *Eur. Phys. J. B* **2013**, *86*, 289. [CrossRef]
38. Leonardos, S.; Piliouras, G.; Spendlove, K. Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26318–26331.
39. Goeree, J.K.; Holt, C.A.; Palfrey, T.R. Regular quantal response equilibrium. *Exp. Econ.* **2005**, *8*, 347–367. [CrossRef]
40. Goeree, J.K.; Holt, C.A.; Palfrey, T.R. Quantal response equilibrium. In *Quantal Response Equilibrium*; Princeton University Press: Hoboken, NJ, USA, 2016.
41. Rilling, J.K.; Sanfey, A.G.; Aronson, J.A.; Nystrom, L.E.; Cohen, J.D. The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* **2004**, *22*, 1694–1703. [CrossRef]
42. Wolpert, D.H.; Harré, M. It can be smart to be dumb. 2008, *Preprint*.
43. Takagishi, H.; Koizumi, M.; Fujii, T.; Schug, J.; Kameshima, S.; Yamagishi, T. The role of cognitive and emotional perspective taking in economic decision making in the ultimatum game. *PLoS ONE* **2014**, *9*, e108462. [CrossRef]
44. Takagishi, H.; Kameshima, S.; Schug, J.; Koizumi, M.; Yamagishi, T. Theory of mind enhances preference for fairness. *J. Exp. Child Psychol.* **2010**, *105*, 130–137. [CrossRef]
45. Lang, H.; DeAngelo, G.; Bongard, M. Theory of Mind and General Intelligence in Dictator and Ultimatum Games. *Games* **2018**, *9*, 16. [CrossRef]
46. Dunbar, R.I. Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* **1992**, *22*, 469–493. [CrossRef]
47. Dunbar, R.I. The social brain hypothesis. *Evol. Anthropol. Issues News Rev. Issues News Rev.* **1998**, *6*, 178–190. [CrossRef]
48. Dunbar, R.I.; Arnaboldi, V.; Conti, M.; Passarella, A. The structure of online social networks mirrors those in the offline world. *Soc. Netw.* **2015**, *43*, 39–47. [CrossRef]
49. Harré, M.S.; Prokopenko, M. The social brain: Scale-invariant layering of Erdős–Rényi networks in small-scale human societies. *J. R. Soc. Interface* **2016**, *13*, 20160044. [CrossRef]
50. Dunbar, R.I.; Shultz, S. Evolution in the social brain. *Science* **2007**, *317*, 1344–1347. [CrossRef] [PubMed]
51. Powell, J.L.; Lewis, P.A.; Dunbar, R.I.; García-Fiñana, M.; Roberts, N. Orbital prefrontal cortex volume correlates with social cognitive competence. *Neuropsychologia* **2010**, *48*, 3554–3562. [CrossRef]
52. Stiller, J.; Dunbar, R.I. Perspective-taking and memory capacity predict social network size. *Soc. Netw.* **2007**, *29*, 93–104. [CrossRef]
53. Lewis, P.A.; Rezaie, R.; Brown, R.; Roberts, N.; Dunbar, R.I. Ventromedial prefrontal volume predicts understanding of others and social network size. *Neuroimage* **2011**, *57*, 1624–1629. [CrossRef]
54. Harré, M.S. Information theory for agents in artificial intelligence, psychology, and economics. *Entropy* **2021**, *23*, 310. [CrossRef]
55. Ert, E.; Erev, I.; Roth, A.E. A choice prediction competition for social preferences in simple extensive form games: An introduction. *Games* **2011**, *2*, 257–276. [CrossRef]
56. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
57. Connors, M.H.; Burns, B.D.; Campitelli, G. Expertise in complex decision making: The role of search in chess 70 years after de Groot. *Cogn. Sci.* **2011**, *35*, 1567–1579. [CrossRef] [PubMed]
58. Ericsson, K.A. Superior Working Memory in Experts. 2018. Available online: https://www.cambridge.org/core/books/abs/cambridge-handbook-of-expertise-and-expert-performance/superior-working-memory-in-experts/8979912B089C15FC7049AC46F940D012 (accessed on 29 April 2022).

59. Gobet, F.; Charness, N. Expertise in Chess. 2018. Available online: https://psycnet.apa.org/record/2006-10094-030 (accessed on 29 April 2022).
60. Harré, M.; Snyder, A. Intuitive expertise and perceptual templates. *Minds Mach.* **2012**, *22*, 167–182. [CrossRef]
61. Harré, M.; Bossomaier, T.; Snyder, A. The perceptual cues that reshape expert reasoning. *Sci. Rep.* **2012**, *2*, 502. [CrossRef] [PubMed]