Title

Economic probes of mental function and the extraction of computational phenotypes

Kenneth T. Kishida[1] and P. Read Montague[1,2]

1. Virginia Tech Carilion Research Institute & Department of Physics, Virginia Tech

2. Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, WC1N 3BG

Abstract

Economic games are now routinely used to characterize human cognition across multiple dimensions. These games allow for effective computational modeling of mental function because they typically come equipped with notions of optimal play. These optimal solutions provide quantitatively prescribed target functions that can be tracked throughout a neuroimaging experiment and thus open up the possibility for new ways to characterize normal cognition. Here we show early results using the multi-round trust game to generate new ways to characterize mental dysfunction and further show how economic games might provide a useful and novel characterization of psychopathology. In this paper we support these claims with early work using interactive games to probe autism spectrum disorder, attention deficit disorder, and borderline personality disorder. Lastly, we discuss how such game-theoretic probes could produce novel bases for representing healthy cognition and thus provide a way to produce predictive computational models of mental function.

Article

Theoretical approaches to the understanding of human decision-making (Von Neumann, Morgenstern, Rubinstein, & Kuhn, 2007) have provided an excellent framework for ongoing empirical investigations, which measure actual human behavior against theoretically optimal actions (Camerer, 2003). The ability to measure brain responses, particularly with the use of functional magnetic resonance imaging (fMRI, (Ogawa, Lee, Kay, & Tank, 1990; Ogawa, Lee, Nayak, & Glynn, 1990)), associated with these behaviors has lead the development of biological investigations into the relationship between human biology and (ir)rational decision making (Loewenstein, Rick, & Cohen, 2008; Montague & Berns, 2002). The early revelation that humans do not always act in accord with economic theory and the ability to measure brain responses associated with these decisions are beginning to inform and reshape economic theories about human decision making (Camerer, 2003; Loewenstein et al., 2008). Additionally these developments have the potential of generating a whole new perspective on the biological bases of human cognition and decision making by providing a novel entry point for the investigation and discovery of genetic architecture that may bias human behavior.

Biology's guiding theoretical foundation remains to be the evolutionary principle that organisms that we observe in nature are derived from the forces of natural selection on heritable traits (Darwin, 1937). Typically the traits discussed are visually identified morphological features, however these traits aren't the only kind that ought to be sensitive to selective forces. Indeed man's best friend,

Canis lupus familiaris, is a stunning example of the variety of traits that can be developed through selective breeding for both morphological and personality traits (Spady & Ostrander, 2008). The notion that styles of decision making may be heritable is not new to academic interests in game theory and biology (Smith, 1982). The concept of a game strategy as a heritatble phenotype was initially developed by Maynard Smith (Smith, 1982) and was used to develop the concept of an evolutionarily stable solution. Wherein an evolutionary stable solution is an optimal one in that any alternative 'mutant' strategy will not be able to invade it. In general, game theory provides a powerful framework for studying socially interacting agents where the strategies employed are guided by various concepts of optimal play. This framework proposes a natural landscape for the application of computational principles to describe otherwise qualitative features of human experience like fairness and trust. The use of these mathematical depictions of human behavior opens the door to new perspectives from which new dimensions of personality (i.e., styles of decision making) and their biological correlates may emerge. The use of these quantitative depictions of important decision making variables will be useful for characterizing normal and dysfunctional human cognition (Kishida, King-Casas, & Montague, 2010) and can provide a relevant basis for identifying biological bases for human cognition at multiple levels including neurobiological and genetic systems.

In this memorial tribute to John Dickhaut, we focus on the use and development of the multi-round trust game (Chiu et al., 2008; B. King-Casas et al., 2008; B. King-Casas et al., 2005; Tomlin et al., 2006), which was derived from his and

colleagues development of the single round version of the game (Berg, Dickhaut, & McCabe, 1995). We show early results from the marriage of the trust game, human neuroimaging, and quantitative depictions of qualities such as trust and fairness as applied to psychopathologies such as autism spectrum disorder, attention deficit hyperactivity disorder, and borderline personality disorder. These early results point to the ability to characterize new dimensions of human behavior and associated neural processes as expressed in strategic game play.

The single round trust game (Berg et al., 1995)

Berg, Dickhaut, and McCabe employed a single round investment game to investigate trust during economic exchange (Berg et al., 1995). In the single round trust game two players engage anonymously; there is an “investor” (first-mover) and a “trustee” (responder); the investor is endowed with $10 and decides an amount to share with their partner; the sent amount (i.e., “investment”) is tripled on its way to the trustee; the trustee then decides how much, if any, to reciprocate to the investor. In the execution of this game the signals transmitted between the players is restricted to the money sent back and forth. As Berg et al., point out the Nash equilibrium for this game is for no money to initially be sent by the investor since a rational and selfish trustee will keep any money sent their way, thus to maximize ones earnings the selfish investor ought to keep everything. Contrary to this prediction, trust (money sent to the trustee) is observed as is reciprocation (money sent back to the investor) and the authors conclude that trust is likely a “behavioral primitive”, which can be predicted by

evolutionary models (Berg et al., 1995) that maximize long-term genetic fitness over short-term gains through selfish behavior. This interpretation of their results is drawn in contrast to games with repeat interactions where trust can be learned or may show varying degrees of stability. An important point about their conclusion is the notion of a behavioral primitive and its relationship to evolutionary constraints; by expressing trust in a single interaction the results suggest that people carry around within them a bias towards trust and reciprocity. The authors do not propose in detail where such a bias may be stored; however, from a neurobiological perspective this bias must be engendered in the neural architecture both structurally and functionally and can be measured using the right tools. Reducing the biology further suggests that the relationship of this behavioral primitive and the evolutionary model may be more than just theoretical, and recent findings suggest a genetic basis for the behavior expressed in this version of the game (Cesarini et al., 2008; Cesarini, Dawes, Johannesson, Lichtenstein, & Wallace, 2009).

The single round trust game is also used by Berg et al., to investigate a “social history” manipulation wherein anonymous and naïve players are provided information about how previous participants played this game; this relatively mild manipulation was observed to increase trust (Berg et al., 1995) suggesting that learning mechanisms and narratives that modulate expectations are also important in determining the expressed strategies. Recent investigations employing fMRI have begun to investigate neural responses associated with the behavioral gestures exchanged within a mulit-round version of the trust game

(Chiu et al., 2008; B. King-Casas et al., 2008; B. King-Casas et al., 2005; Tomlin et al., 2006). Additionally the use of the multi-round trust game and fMRI has been used to investigate neurobehavioral responses in populations characterized by clinically abnormal social behavior including participants diagnosed with autism spectrum disorder (Chiu et al., 2008) and borderline personality disorder (B. King-Casas et al., 2008). These studies demonstrate early developments in using game theory and computational approaches for understanding mental disorders (Kishida et al., 2010), which are believed to be strongly influenced by genetic predispositions. Below, we discuss the multi-round trust game, early neurobehavioral findings, and the direction this work may take in order to determine the degree to which game theoretic parameters may be used to characterize heritable quantitative phenotypes.

Multi-round trust game and computational models of learning

The multi-round trust game (B. King-Casas et al., 2005; Tomlin et al., 2006) allows the investigation of signals associated with iterated social exchange, including agent detection (Chiu et al., 2008; Tomlin et al., 2006), learning, and the development and expression of expectations (B. King-Casas et al., 2005). The initial development of the single round trust game (Berg et al., 1995) intended to reduce the effects of knowledge and reputation in order to examine the underlying bias regarding trust and selfish decision making, whereas the multi-round version aims to study these processes while eavesdropping on the underlying neural processes. Like the single round version two players engage anonymously; there is an “investor” (first-mover) and a “trustee” (responder); the

first round is implemented in the same manner as the single-round version, however, subjects know that they will engage in a total of ten iterative rounds with the same partner (Figure 1). This manipulation allows the study of signals sent between participants that know there will be feedback and a chance to respond to that feedback. It also allows the investigation of the modulation and development of internal models about the intentions and beliefs expressed between the two agents. Along these lines King-Casas and colleagues measured brain responses during the multi-round trust game using functional magnetic resonance imaging and identified brain responses consistent with reinforcement learning signals previously associated with dopaminergic neural activity (B. King-Casas et al., 2005).

King-Casas et al. identified these brain responses by taking advantage of the ability to quantify and computationally model the “social gestures” in the context of the game. Expressions of increases or decreases in trust are captured by changes in the values sent to ones’ partner from one round to the next (Figure 2). In early rounds, increases in trust by the trustee (i.e., increases in reciprocity) are preceded by an increase in a response (black trace, top right panel of Figure 2) in the striatum (Figure 2, left inset) following revelation of the amount of money sent from ones’ partner. This response is consistent with reward-related processing of a social gesture leading to increased reciprocation of trust. On the other hand, subsequent decreases in trust are not preceded by an increase in striatal responses (red trace, top right panel of Figure 2). Interestingly, in later rounds the striatal response becomes anticipatory and responds to the earliest

phase of the trial where a positive signal can be predicted (black trace, bottom left panel of Figure 2). Here the trustee brain may be predicting a positive signal and when the expectation is met an increase in trust is delivered. These results are consistent with reputation formation and the development of positive expectations of trust between the two partners. These results also suggest something more fundamental; the pattern of activity observed in the striatum matches very closely with learning dynamics previously observed in the dopaminergic system in non-human primates engaged in a simple Pavlovian learning paradigm (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). The computational depiction of these simple learning signals predicts the observed temporal shift in the response pattern observed in the dopaminergic system and those observed in the King-casas et al social exchange with brain imaging study (B. King-Casas et al., 2005). Further work suggests that the possibility that the dopaminergic system may serve as a common valuation system during learning and decision making in a wide range of valuation scenarios (Behrens, Hunt, Woolrich, & Rushworth, 2008; Behrens, Hunt, & Rushworth, 2009; Chiu, Lohrenz, & Montague, 2008; Fliessbach et al., 2007; Izuma, Saito, & Sadato, 2008; Kishida et al., 2011; Kishida, Yang, Quartz, Quartz, & Montague, in press; Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009; Lohrenz, McCabe, Camerer, & Montague, 2007; McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Pagnoni, Zink, Montague, & Berns, 2002; Seymour et al., 2004; Zink et al., 2008).

Biologically interesting participants can be anonymously interchanged in the trustee role to investigate how these individuals modulate the iterated dynamic exchange. This kind of manipulation has been carried out to investigate a range of psychopathologies including participants diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, borderline personality disorder, and major depression (Chiu et al., 2008; B. King-Casas et al., 2008; Koshelev, Lohrenz, Vannucci, & Montague, 2010).

Autism spectrum disorder, attention deficit hyperactivity disorder, and borderline personality disorder in the multi-round trust game

These are still early days in the investigation of psychopathologies using computational approaches in game theoretic settings and neuroimaging (Kishida et al., 2010). However, there are already promising developments where different categories of psychopathology are showing differentiating strategies in game behavior. If one allows brain responses to be included in a more general class of expressible behavior then increased dimensionality may be provided to the problem of classifying dysfunctional mental processes. Recent successes of the application of game theory to mental disorders include investigations into autism spectrum disorder (Figure 3 adapted from (Chiu et al., 2008)) and borderline personality disorder (B. King-Casas et al., 2008). Chiu and colleagues used the multi-round trust game and hyperscanning (Montague et al., 2002) to investigate social exchange in individuals diagnosed with autism spectrum disorder (ASD). The participants diagnosed with ASD were assigned to the trustee role and compared to age-matched participant in the trustee role. These participants were

relatively high functioning (as assessed by an estimate of their IQ) and repaid their investor quite similarly to age-matched controls in the multi-round trust game (Figure 3A from (Chiu et al., 2008)). A challenging feature of the trust game is that participants must either possess or develop an accurate model of their partner in order to maximize their returns. Chiu et al showed that a previously described agent-specific response in the cingulate cortex (Tomlin et al., 2006) was diminished in the ASD cohort (Figure 3B from (Chiu et al., 2008)). Specifically, a spatial pattern of activity dubbed the “cingulate self response”, which was observed in contrast to the “cingulate other response” (Tomlin et al., 2006), was shown to be diminished proportional to the participants symptom severity (Figure 3B from (Chiu et al., 2008)). Further work suggests that the cingulate self response pattern, which was only observed during real social exchange (versus simulated game play), is associated with perspective-taking (Chiu et al., 2008).

The relatively cooperative behavior typically observed in the trust game suggests that players share norms about fairness in these kinds of exchanges and reciprocation of trust appears to be normal behavior. This normative observation suggests that some psychopathologies may be more or less sensitive to signals and calculations of fairness and equitable distributions. We investigated whether the ASD participants experienced more of less equitable exchange during the multi-round trust game (Figure 4). Inequity is simply calculated as a deviation from equivalence (i.e., when either the investor or trustee ends up with more than their partner at the end of a round, see Figure 4 for more detail). ASD

participants experienced significantly more inequitable exchange compared to age-matched controls, participants diagnosed with attention deficit hyperactivity disorder, or borderline personality disorder (Figure 4A: $p < 0.001$, two-tailed t-test). A round-by-round analysis (Figure 4B) shows that the ASD participants experienced more inequitable exchanges in the latter half of the game (Figure 4B: rounds 6, 7, and 8: $p < 0.05$, two tailed t-test).

Another normative behavior in the trust game includes relatively high initial investments; these signals may be derived from an initially shared expectation of cooperation thus resulting in expectations of high investment. Deviations from these expectations may result in social exchange dynamics that lead to the break down in trust; problems can also arise when partners do not share the same model of what pro-cooperative signals look like. King-Casas et al studied trust game behavior and the associated neural responses in individuals diagnosed with borderline personality disorder (BPD) (B. King-Casas et al., 2008). Individuals diagnosed with BPD demonstrate pervasive instability of interpersonal relationships, self-image, and affect, which begins early in adult life (*American psychiatric association: Diagnostic and statistical manual of mental disorders, fourth edition, text revision*2000). In the King-Casas et al study, the control population (trustee role) showed brain responses in the insula cortex that correlated with diminishing investments from the investor participants. This is consistent with the insula detecting norm violations in a range of experimental paradigms (Montague & Lohrenz, 2007). Interestingly participants diagnosed with BPD showed no parametric relationship between the size of the offer and their

insula response, rather their insula was consistently activated by all offer sizes (B. King-Casas et al., 2008). This neural strategy was associated with an unwillingness to cooperate over the multi-round interactions. Additionally, investors (unknowingly) playing with a BPD trustee were nearly half as likely to send gestures consistent with coaxing behavior (B. King-Casas et al., 2008). Coaxing is typically seen in pairs of healthy participants in response to low offers and is considered to be an attempt to draw the partner into a cooperative mode (B. King-Casas et al., 2008); the data suggest that the relatively low bandwidth signaling afforded by the trust game setup provided enough information to the investor to alter the participants' behavior in a meaningful way. Interestingly, the participant pairs consisting of a trustee diagnosed with BPD do not show deviations form fair splits (Figure 4A).

Can game-theoretic probes shed new light for phenotyping humans?

Game theory has provided a powerful framework for considering how an idealized agent ought to behave in highly structured social exchange games (Von Neumann et al., 2007). Indeed, the framing of experiments by game theory has already provided much insight into the principles that accurately characterize human decision making, but more interesting are the cases where human behavior deviates from economic theory (Camerer, 2003). Additionally, game theoretic approaches have provided interesting insights into the evolution of non-human organisms and the strategies they employ (Smith, 1982; Smith & Harper, 2003). Human neuroimaging results suggest that the presumed valuation machinery in human brains does not only respond to monetary gains and losses,

which are an important guidance signal in economic theory; the data suggest that primary rewards and social status (in the absence of monetary gains and losses) can also effect the same neural machinery. This points to a biologically guided valuation system where money is likely a proxy or a cue for something more fundamental. This idea is consistent with the approach taken by evolutionary biologists where genetic fitness is the guiding principle; strategies (i.e., phenotypes) are selected that maximize genetic fitness. Certainly humans are not exempt from the pressures of natural selection. Our genome and the biological processes it dictates have resulted from generations of successful strategies defined by increased success in reproducing and surviving. Models that describe evolutionary stable strategies within the context of games like the trust games (Berg et al., 1995; B. King-Casas et al., 2005) can provide us clues regarding the status of human evolution and the direction our species in developing; deviations from evolutionary stability may suggest a kind of distance from equilibrium and the evolutionary age of our cognitive capacities. This in turn would provide us theoretical clues regarding the dynamics of underlying genetic architecture.

We have presented quantitative neural and behavioral results that identify specific deviations from normal behavior in the context of the multi-round trust game. The diminished cingulate response in ASD subjects is specific to this population (BPD patients do not show a deviation in this response, data not shown). Also, the behavioral results shown in Figure 4 demonstrate specificity in the response to inequitable gestures; only the ASD participants show increased

inequitable exchanges. Our early results in this domain suggest that games and the expression of strategies in human populations can be fruitfully explored and may lead to the characterization of evolutionarily stable strategies in human behavior. The concept of an evolutionary stable strategy (Smith, 1982) suggests theory based strategies for identifying heritable “decision-making traits”, which take a parametric form in the context of games and their mathematical depictions. These traits would be exposed in experiments that sample strategic decision-making in large samples of human populations. Measuring the distribution of any quantitative trait within these games will begin to characterize what would be considered normal/healthy human cognition within these dimensions and would set the stage for identifying subpopulations of aberrant decision-making phenotypes. Qualitatively the DSM-IV criteria for mental disorders achieve this, but without the quantitative rigor or objective threshold criteria that game theoretic approaches promise and that genetic discoveries may require. Characterizing the multidimensionality of mental disorders in a naturally quantitative framework will allow computational tools new to the investigation of mental health to explore a powerful and novel perspective on an old problem and may provide insight into how the high-level psychological depictions may be reduced. The reduction of human personality and subjective experience as expressed in human decision-making behavior in the context of games is an exciting and relatively unexplored direction for investigation into the biological basis for human psychology.

References

American psychiatric association: Diagnostic and statistical manual of mental disorders, fourth edition, text revision (2000). (Fourth, Text Revision ed.). Washington, DC: American Psychiatric Association.

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science, 324*(5931), 1160.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*(7219), 245.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction* Russell Sage Foundation New York.

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of Sciences of the United States of America, 105*(10), 3721-3726. doi:10.1073/pnas.0710069105

Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Experimental game theory and behavior genetics. *Annals of the New York Academy of Sciences, 1167*, 66-75. doi:10.1111/j.1749-6632.2009.04505.x

Chiu, P. H., Kayali, M. A., Kishida, K. T., Tomlin, D., Klinger, L. G., Klinger, M. R., & Montague, P. R. (2008). Self responses along cingulate cortex reveal

quantitative neural phenotype for high-functioning autism. *Neuron, 57*(3), 463-473. doi:10.1016/j.neuron.2007.12.020

Chiu, P. H., Lohrenz, T. M., & Montague, P. R. (2008). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuroscience, 11*(4), 514-520. doi:10.1038/nn2067

Darwin, C. (1937). *The origin of species*

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E., & Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science, 318*(5854), 1305.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*(2), 284-294.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science, 321*(5890), 806.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science (New York, N.Y.), 308*(5718), 78-83. doi:10.1126/science.1108062

Kishida, K. T., King-Casas, B., & Montague, P. R. (2010). Neuroeconomic approaches to mental disorders. *Neuron, 67*(4), 543-554.

Kishida, K. T., Sandberg, S. G., Lohrenz, T., Comair, Y. G., Sáez, I., Phillips, P. E. M., & Montague, P. R. (2011). Sub-second dopamine detection in human striatum. *PloS One, 6*(8), e23291.

Kishida, K. T., Yang, D., Quartz, K., Quartz, S., & Montague, P. R. (in press). Implicit signals in small group settings and their impact on the expression of cognitive capacity and associated brain responses. *Philosophical Transactions of the Royal Society of London.B, Biological Sciences,*

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61*(1), 140-151.

Koshelev, M., Lohrenz, T., Vannucci, M., & Montague, P. R. (2010). Biosensor approach to psychopathology classification. *PLoS Computational Biology, 6*(10), e1000966.

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annu.Rev.Psychol., 59*, 647-672.

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences of the United States of America, 104*(22), 9493-9498. doi:10.1073/pnas.0608842104

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38*(2), 339-346.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron, 36*(2), 265-284.

Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., . . . Fisher, R. E. (2002). Hyperscanning: Simultaneous fMRI during linked social interactions. *NeuroImage, 16*(4), 1159-1164.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 16*(5), 1936-1947.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron, 56*(1), 14-18. doi:10.1016/j.neuron.2007.09.020

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38*(2), 329-337.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America, 87*(24), 9868-9872.

Ogawa, S., Lee, T. M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine : Official Journal of the Society of Magnetic*

Resonance in Medicine / Society of Magnetic Resonance in Medicine, 14(1), 68-78.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience, 5*(2), 97-98.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., . . . Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature, 429*(6992), 664-667.

Smith, J. M. (1982). *Evolution and the theory of games* Cambridge Univ Pr.

Smith, J. M., & Harper, D. (2003). *Animal signals* Oxford University Press, USA.

Spady, T. C., & Ostrander, E. A. (2008). Canine behavioral genetics: Pointing out the phenotypes and herding up the genes. *The American Journal of Human Genetics, 82*(1), 10-18.

Tomlin, D., Kayali, M. A., King-Casas, B., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2006). Agent-specific responses in the cingulate cortex during economic exchanges. *Science (New York, N.Y.), 312*(5776), 1047-1050. doi:10.1126/science.1125596

Von Neumann, J., Morgenstern, O., Rubinstein, A., & Kuhn, H. W. (2007). *Theory of games and economic behavior* (16th ed.) Princeton Univ Pr.

Zink, C. F., Tong, Y., Chen, Q., Bassett, D. S., Stein, J. L., & Meyer-Lindenberg, A. (2008). Know your place: Neural processing of social hierarchy in humans. *Neuron, 58*(2), 273-283.

Figure Legends

Figure 1. Multi-round trust game probes social exchange in known psychopathological categories.

The multi-round trust game (B. King-Casas et al., 2005; Tomlin et al., 2006) is a repeated interaction (10 - round) version of the single round trust game (Berg et al., 1995). An “investor” is given an initial endowment and is to choose how much to share with his/her partner. This “investment, *i* “ is tripled on its way to the “trustee”. The trustee then chooses how much of “3*i*” to send back to the investor. The total points each player earns in a single round is placed into a “bank” and the game is repeated for a total of ten rounds. This deviation from the single round version allows the observation and measurement of reputation formation and learning signals embedded in this simple interaction. The multi-round trust game has been used to probe social exchange in a number of ”patient” categories classified by DSM-IV criteria including: autism spectrum disorder, borderline personality disorder, and attention deficit hyperactivity disorder.

Figure 2. Hyperscanning during two-person trust game reveals the development of signals for reputation formation. (figure adapted from (B. King-Casas et al., 2005)**)**

Left: Brain responses in the trustees’ brain to “benevolent” investor behavior. Statistical parametric map showing significant activation in the bilateral head of the caudate nucleus in the trustees’ brain for “better than expected” behavioral gestures from the investor (n = 125 gestures).

Right: Neural correlates of reputation building. Blood-oxygen-level-dependent (BOLD) responses from the regions defined in Figure 2A; time series of the BOLD response is time locked to the "investment" revelation, but separated according to what the trustees' next decision. Black: future increase in trust; red: future decrease in trust. In early rounds (top rows) a significant increase in the BOLD response in the caudate follows investment revelations that lead to the trustee increasing their trust (black trace). This signal undergoes a temporal transfer in later rounds (bottom rows) to just prior to investment revelation, which suggests that the trustee brain is anticipating trustworthy investments from the investor before they are revealed.

Figure 3. Multi-round trust game reveals diminished cingulate response in participants diagnosed with autism spectrum disorders (adapted from Chiu et al 2008 (Chiu et al., 2008)**).**

A. Average Trustee repayment ratio round-by-round. The repayment ratios are not significantly different round-by-round in ASD participants compared to controls.

B. Diminished cingulate response pattern during "self phase" of the iterated multi-round trust game. Left: heat maps showing spatial pattern of activity indicative of self- and other-responses during the multi-round trust game (Tomlin et al., 2006), where the cingulate self-response is revealed to be specifically diminished in individuals diagnosed with ASD (see response labeled with white asterisk). Right: the magnitude of signal change in the middle portions of the cingulate cortex during the self-response phase of the task show

significant correlation with the assessment of ASD symptom severity (Chiu et al., 2008) (open circles: ADI communication subscale, r = -.69, p = -.012; light blue filled circles: ADI social subscale, r = -.70, p = .011; dark blue filled circles: ADI total score, r = -.73, p = .007).

Figure 4. ASD participants experience greater inequity, round-by-round, than other participant cohorts.

A. Participant pairs consisting of ASD trustees experience greater average inequity over all rounds played. For a given round, "*t*", inequity (inequity(*i,r,t*) = $e - 4i_t + 6i_t r_t$) is calculated as a function of the endowment, "*e*", investment, "*I",* sent by the investor and the repayment, "*r*", sent by the trustee. *i* and *r* are expressed as a fraction of each participants holdings in that round respectively and the range of inequity ratios is -3 to +3. An inequity ratio of 0 indicates that the points in a given round are evenly split between the two players; a positive inequity ratio indicates that the investor received more in the exchange than the trustee; and negative values indicate the opposite (trustee > investor). Age-matched controls (compared to the ASD participants), participants diagnosed with ADHD, and participants diagnosed with BPD experience equitable exchanges on average, but ASD participants experience greater inequity in favor of the investor (paired asterisks indicate significant difference $p < 0.001$, two-tailed t-test).

B. Round-by-round assessment of inequity in participants diagnosed with ASD. Inequity is calculated in the same manner as in panel A; here we compare the mean inequity experienced during each round of the multi-round trust game.

In early rounds the participant pairs with the ASD participant in the trustee role experience similar equitable exchanges compared to age-matched controls. In later rounds, the inequitable distribution goes in favor of the investor only in pairs consisting of trustees diagnosed with ASD (asterisks indicate significant difference, $p < 0.05$, two-sample t-test).

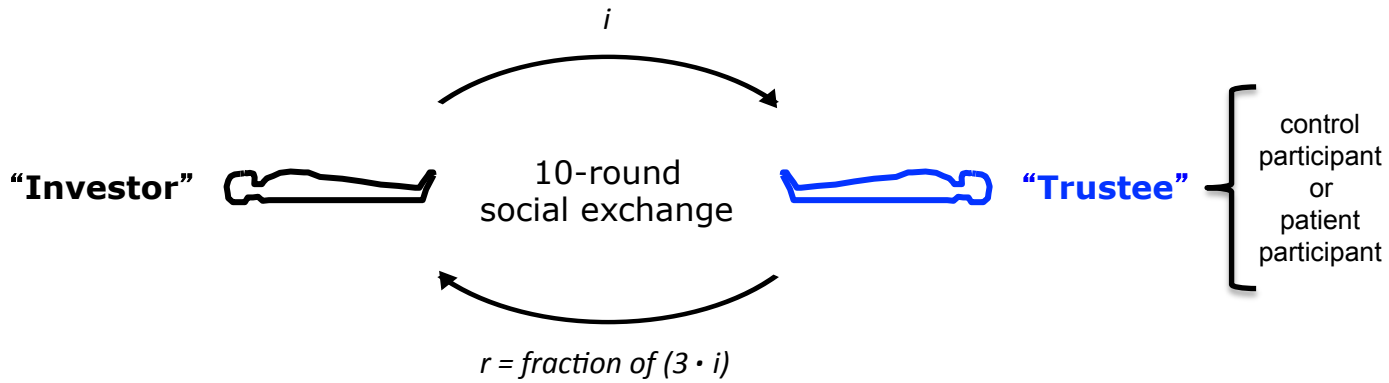

i
"Investor"
10-round
social exchange
"Trustee"
control
participant
or
patient
participant
r = fraction of (3 • i)

Figure 1

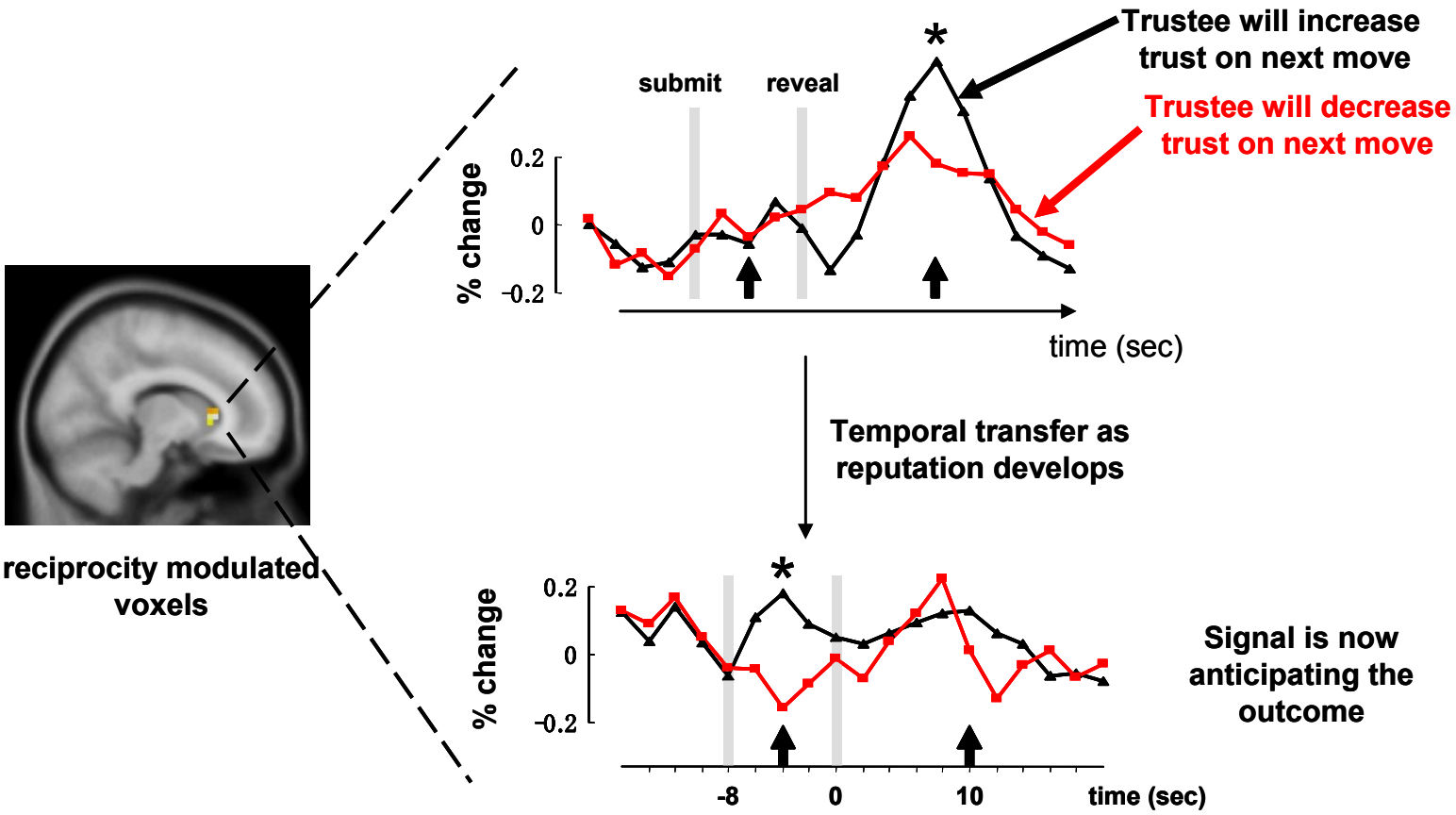
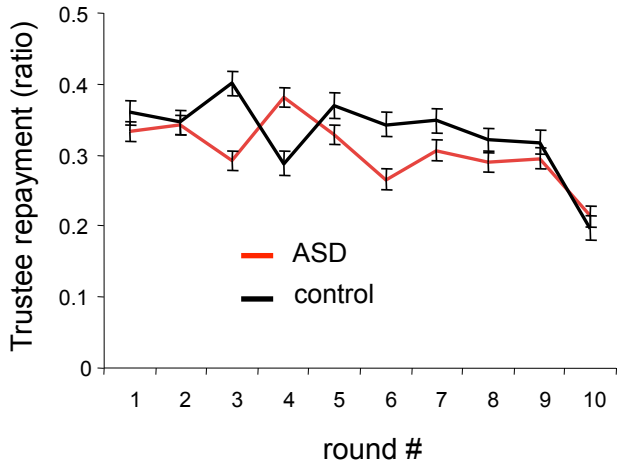
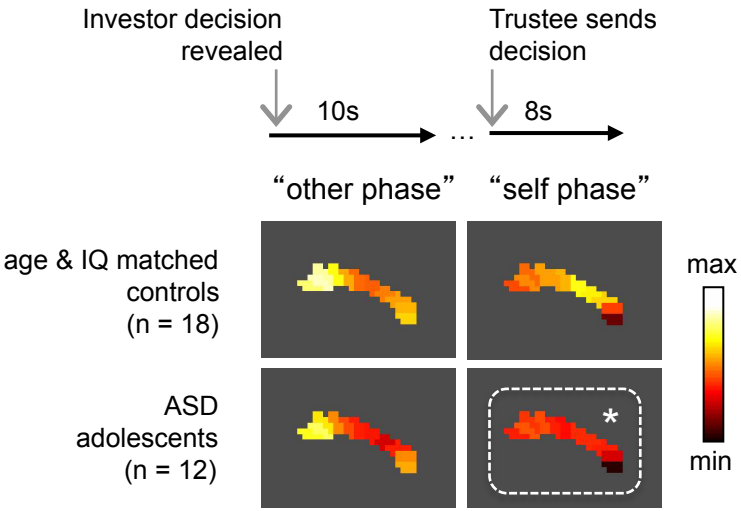

Trustee will increase trust on next move
Trustee will decrease trust on next move
submit
reveal
*
% change
0.2
0
-0.2
time (sec)
Temporal transfer as reputation develops
reciprocity modulated voxels
*
% change
0.2
0
-0.2
-8
0
10
time (sec)
Signal is now anticipating the outcome

Figure 2

A


Trustee repayment (ratio)
0.5
0.4
0.3
0.2
0.1
0
ASD
control
1
2
3
4
5
6
7
8
9
10
round #

B


Investor decision revealed
Trustee sends decision
10s
8s
…
“other phase”
“self phase”
age & IQ matched controls (n = 18)
ASD adolescents (n = 12)
max
min
*

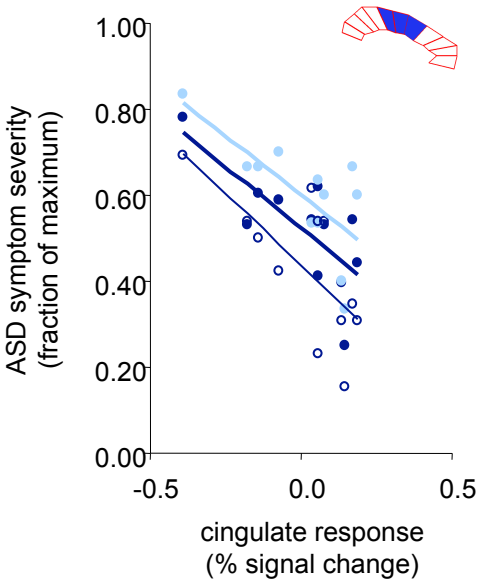

ASD symptom severity (fraction of maximum)
1.00
0.80
0.60
0.40
0.20
0.00
-0.5
0.0
0.5
cingulate response (% signal change)

Figure 3

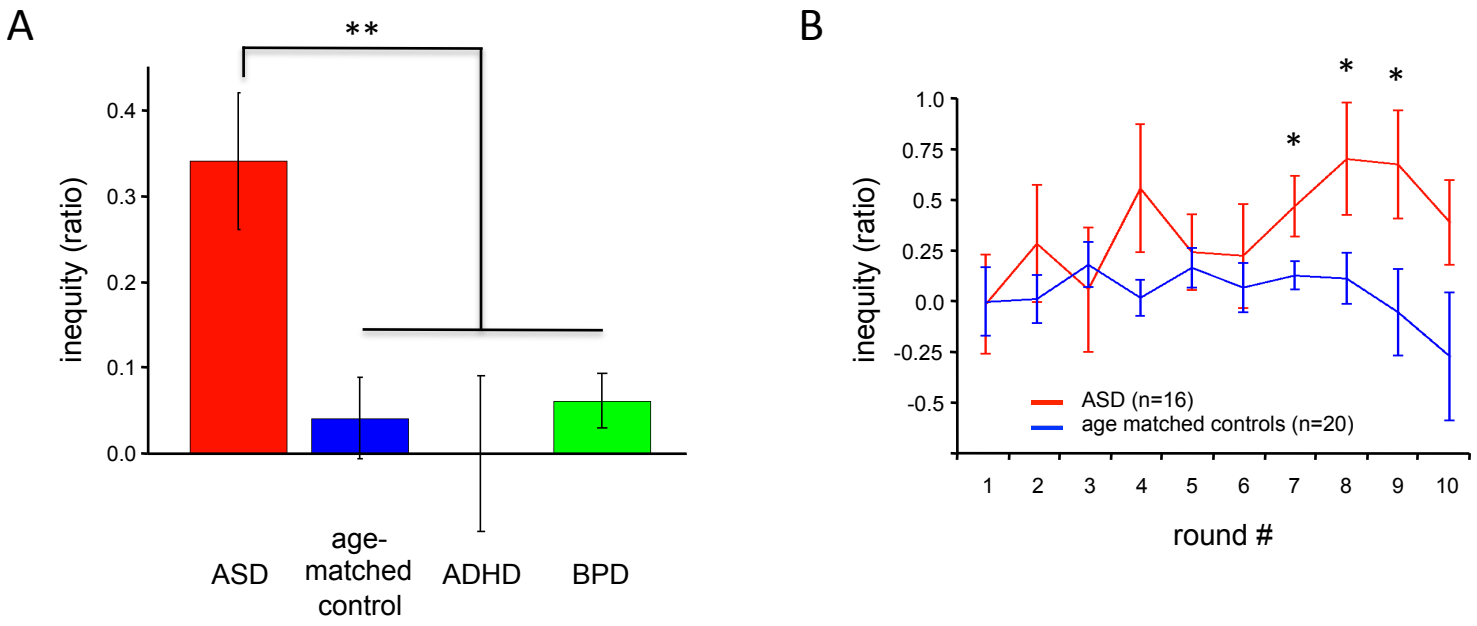

A
**
inequity (ratio)
0.4
0.3
0.2
0.1
0.0
ASD
age-matched control
ADHD
BPD
B
*
*
*
1.0
0.75
0.5
0.25
0.0
-0.25
-0.5
inequity (ratio)
ASD (n=16)
age matched controls (n=20)
1
2
3
4
5
6
7
8
9
10
round #

Figure 4