Review of Trust and Machine Ethics Research: Towards A Bio-Inspired Computational Model of Ethical Trust (CMET)

Hock Chuan Lim
ITEE, UNSW@ADFA
Canberra, Australia
hc.lim@adfa.edu.au

Rob Stocker
ITEE, UNSW@ADFA
Canberra, Australia
r.stocker@adfa.edu.au

Henry Larkin
ITEE, UNSW@ADFA
Canberra, Australia
h.larkin@adfa.edu.au

ABSTRACT

Recent advances in the fields of robotics, cyborg development, moral psychology, trust, multi agent-based systems and socionics have raised the need for a better understanding of ethics, moral reasoning, judgment and decision-making within the system of man and machines. Here we seek to understand key research questions concerning the interplay of ethical trust at the individual level and the social moral norms at the collective end. We review salient works in the fields of trust and machine ethics research, underscore the importance and the need for a deeper understanding of ethical trust at the individual level and the development of collective social moral norms. Drawing upon the recent findings from neural sciences on mirror-neuron system (MNS) and social cognition, we present a bio-inspired Computational Model of Ethical Trust (CMET) to allow investigations of the interplay of ethical trust and social moral norms.

Keywords

Ethical Trust,Neural Network, Mirror Neuron System (MNS)

1. INTRODUCTION

Significant advances in research from the fields of robotics, neural sciences [28, 71, 86], cyborg development [90], moral psychology, multi agent-based systems (MAS) [21] and socionics [54] have spurred a need in the study of the ethical dimension of machine or Machine Ethics. Recent papers by key researchers from an interdisciplinary field of computer science, philosophy, psychology and sociology have attested to the need and importance of a better understanding of this new field [3, 4, 5, 66] while others [50, 64] have argued for the potentials of neural processes for a deeper understanding of self and other social cognitive behaviors.

Within these studies, issues of moral reasoning at the individual level and moral norms development at the social level are of particular interest. A central theme that complements these interests is the notion of ethical trust – a form of trust that is separate from the more traditionally recognized social and cognitive dimensions. We contend that ethical trust is more than just rational decision-making based on options, values or cost. It is about thinking of others, and for others based on accepted ethics and moral norms. Through a review of selected important research in trust and machine ethics, we show why this ethical dimension of trust is significant and how we can (through the use of core computer-based investigative techniques inspired by neural agent-based processes) contribute to a better understanding of the interplay of ethical trust at the individual level and social moral norms at the collective level.

The key research questions that this paper seeks to understand are:

(1) How can computer-based investigative techniques (for example agent-based modeling and simulation) assist in understanding complex topics of ethical trust, particularly in the domain of moral reasoning, moral judgment and moral actions;

(2) Do moral judgments always translate to moral actions at the individual level? [10] suggests that this may not always be so.

(3) How do moral judgments and actions at the individual level translate to collective moral actions at the group or communal level? Does one's moral action become a collective moral norm for a group?

Section 2 reviews important research on trust and machine ethics. Section 3 outlines the conceptual notion of a bio-inspired Computational Model of Ethical Trust (CMET). A two-tier architecture is proposed and aspects of the model design are discussed. Key features of the CMET are ethical trust reasoning and bio-inspired neural agent-based processes in the evolution of moral norms. Section 4 provides concluding discussions that lead to works including the application of CMET to investigate the interplay of ethical trust and moral norms.

2. RESEARCH IN TRUST AND ETHICS

Here we analyse selected research in trust and ethics.

2.1 Trust

The notion of trust is an interesting one. In fact, according to Stanford Encyclopedia of Philosophy, "trust is both important and dangerous". It is important as trust is an essential element of interactions and is dangerous as trust involved risk taking. Trust can exist in many forms in different contexts [61], for example, it can be a cause as in "

Bionetics '08, November 25-28, 2008, Hyogo, Japan

trust is a needed ...", or it can be an effect as in " trust is seen as ..."; even as an organization as in "A Trust Association ..." or a commodity as in "the Trust fund". Trust has long been the subject of multi-disciplinary attention and the general perception is that trust is something good. Contrary to this general view, back in 1759, Adam Smith provided an early account of a different kind of trust when he reflected that "if there is any society among robbers and murderers, they must at least, according to the trite observation, abstain from robbing and murdering one another" ([84] : p86); [26] went on to add that "we may want *less* cooperation (and trust) rather than more, especially among those who are threatening us" ([26] emphasis in original). This notion that more trust may not be always a positive value had been raised before by [15] when he surveyed the trust literatures. Of course, the surveys were about "Trust organizations or great combination of capitals" as it was called in those days and their impacts on markets notably in the field of economics. Since then, the notion of trust has been studied extensively in many different fields: in economics [35, 93]; in sociology (See [49, 52, 65, 80, 83]); in management as initial, organizational and institutional trust (See [42, 58, 81]) and more recently in computer science (See [9, 17, 32, 57, 62, 69]). Given the highly complex nature of trust, some researchers even go as far as to suggest that trust be suitably limited in context (See [35, 93]).

Two important developments can be traced in trust research. First, we are informed of the different dimensionality of trust, in terms of social [16, 49, 52, 68], affective and cognitive-based trust [41, 59]. This has allowed trust to be link to social capital [23, 46] and the neural science research while at the same time adding interesting perspectives. Second, we are informed of the computational aspects of trust from the research by [55]. This work served as a primer for other computational aspects of trust investigations (See also [43, 44, 56]). We suggest that this trend – that of adding the social and the computational dimensions of trust have lead to greater understanding of the concept.

2.2 Machine Ethics

Ethics like trust has long received the active attention of philosophers ranging from those of pre-Socratic era to those in our current modern day context (See [73, 88, 91]). It is not possible within this limited space to list all their works except suffice to say that they all in their own way addressed and contributed to the growth of the study of morals and ethics. In fact, this field of study is wide and has a rich history especially when one considers both eastern and western traditions and thoughts.

It is interesting to note that morals and ethics like trust are showing a similar trend that we have mentioned in the earlier section. They too are moving towards adding the social and computational dimensions. For instance, recent research in the traditions of "affective revolution" by [37] suggested a fourth principle to guide research as in "morality is more than harm and fairness" and suggest that morality has a social flavour. In addition, a new field of machine ethics research (See [3, 66]) is working towards putting computational ethics into machines.

In this new direction of machine ethics research, a number of novel studies aimed at ways of computing ethics have been initiated. These studies include: [18, 19] used game-theoretic approach; [14, 87] used formal logic of deontic, epistemic and action logic; [63] adopted a case-based approach; [33, 34] used neural network approach; [75] used Kantian rule-based ethical theories; [27] used answer set programming to model ethical rules; [4, 6] used inductive logic programming and theories of multiple prima facie duties; [74] employed prospective logic, a model based on abductive reasoning and a look-ahead feature; and finally [8] addressed ethical controllers in robotics; [12] addressed the ethics of autonomous military robots; and [89] touched on implementing moral decision making facilities in computers and robots. Note that these projects focused more on ethics from the western philosophical perspective. Still, with this trend of having a social and computational dimension of ethics and morality, we can expect greater interactions of man and machines.

The need to better understand trust from its ethical dimension is necessary. We see the need for ethical trust to balance the uneven expansion of technological advances such as in robotics and Cyborg development. The two terms of trust and ethics have been mentioned in passing by a few authors. The most interesting observations can be drawn from [40] that trust is based on moral duty; [92] that ethics and management needs optimal trust; and [83] of trust and morals within his theories on secret and the secret society. Modern society with greater interactions in the system of man and machines will need both ethics and trust.

2.3 Significance of Ethical Trust

The review of researches in trust and machine ethics allows us to group the need for ethical trust into two perspectives:

1. Man to Machines. In this perspective, trust is essential to ensure proper social interactions and order [52] and hence, by the same extension so is ethical trust needed from man to machines. We will need to have the right reason to be able to trust machines that are autonomous and where their actions have ethical implications.

2. Machine to Man. Here we need to define the scope and boundaries of ethical trust with which we want machines to be built and to interact with us.

From these perspectives, a greater understanding of ethical trust will offer us the following benefits:

- Better control of machine autonomy; restraint of machine freedom of choice and behaviour especially in the medical and healthcare domains;
- Better control of cyborg development and experimental development of biological weapons;
- Better awareness of ethics and nano-technology;
- Greater confidence towards pervasive computing technologies and products;
- Greater trust in the use of the Internet;
- Options to investigate social and moral development theories;
- Extend the frontiers of scientific research through better use of agent-based computing techniques;

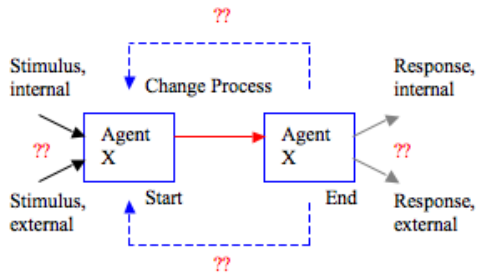

Figure 1: A Simple Agent World Environment

We develop CMET as a model to allow us to investigate conditions of individual ethical trust reasoning and its interplay with social moral norms.

3. COMPUTATIONAL MODEL OF ETHICAL TRUST (CMET)

Here we address the conceptual design and architecture of CMET.

3.1 Preliminaries – Agent State Change

In an agent environment, we observe that an agent can change from one valid state to another. It is possible to view this change as a process made up of signals where certain stimuli serve as the inputs and valid responses are the outputs. This simplified agent world environment is shown with arrows and blocks diagram in Figure 1. A block represents a single agent X. Agent X is any form of valid agency such as a human agent or a software agent. The arrow between the Agents represents a change process. This change process is depicted based on the notion of a signal flow. This process has stimuli as inputs and shows the change of state as a directional signal that flows from the agent X's start state to its end state. In this change process, a change is said to have taken place if and only if a signal is generated and this signal causes a change of agent X's state from a certain start state to a certain end state that are both conceptually measurable. The small arrows on the left represent two types of input or stimulus – internal and external stimuli that initiate the change process. The arrows on the right represent two types of output or response – internal and external response. The dotted arrows depict the possible influence or feedbacks between the responses and the stimuli. The "double question marks" symbol represents the interplay between the various elements.

We next map the ethical dimension with this agent world environment model from Figure 1. The external stimulus is the specific ethical case or dilemma that an agent can encounter. The internal stimulus refers to moral intuition, values and principles that an agent uses to effect a change. We term these internal and external stimuli as the inputs needed for ethical reasoning. The internal response refers to the moral judgment and decisions and the external response refers to the moral conduct or action. They are the outputs from the ethical reasoning process.

From this simplified agent world very little is known about the interplay of the internal response (moral judgment) and that of the external response (moral action). A common and simple way has been to make a simplified assumption – that is, moral judgment and decision naturally leads to moral actions. While this may be true for certain specific cases especially at the individual level, this assumption is too naïve, too weak and inappropriate for a deeper investigation and understanding of the field on moral reasoning. Indeed, it is from this formulation that we have derived the list of research questions that we have outlined in the first section of this paper.

Given that we are not clear about the interplay between moral judgment and moral action, a more useful approach would be to view the moral judgment and moral action from separate but inter-related perspectives and model their interactions. This forms a core design consideration in the formulation of our Computational Model of Ethical Trust (CMET). We model CMET as an expansion of the "arrow of change process" shown in Figure 1, to incorporate the working mechanisms of the change. In addition to modelling ethical reasoning, in our CMET model, we deploy bio-inspired neural mechanisms that are agent-based. The bio-inspirations include the notion of mirror neuron systems (MNS), equipped with agent functionality of "mirroring" and "adaptive learning".

3.2 CMET – A Two-Tier Architecture

Recent neural research suggests that reasoning of social and moral behaviours are related to various areas of the human brain particularly "a functional network of brain regions including the ventromedial prefrontal cortex (VMPFC), orbitofrontal cortex (OFC), the temporal poles, the amygdala, the posterior cingulate cortex (PCC) and the posterior superior temporal sulcus (PSTS)" ([76]:p 33) (For more detailed discussions on neural science research, social and moral behaviors, see also [2, 7, 31, 38, 47, 77]).

These investigations provide intuitions that ethical considerations and trust behaviour exist and that they are important components of our reasoning system. While studies have found that the cognitive and the affective elements mutually affect each other [1, 48] it is likely that they play an important role in the moral reasoning process and it is also highly likely that the situated social context impacts the overall reasoning process (See [20]). In this context, [30] suggested that moral thinking could involve two types of processes a "domain-specific, social-emotional responses and domain-neutral reasoning processes applied in moral contexts".

Hence, we can posit a type of moral thinking. Ethical trust reasoning can be conceived as a form and a process of reasoning involving the ethical or moral consideration of others and a willingness to accept the risk or by exercising a "leap of faith" to arrive at a moral decision and/or action.

Taking recent research findings into consideration, CMET is designed as a two-tier architecture. Research on moral reasoning and process changes suggest that we reason from a dual-level paradigm that of a general-purpose socio-cultural level; and a specific-experience paradigm or a two-tiered cognitive architecture [11]; (See also [36, 72]). In many real world contexts, we are confronted with notions of moral intuitions, moral judgment, moral justifications and moral actions where each of these notions may have different roles and impacts on the reasoning process. An appropriate approach would be to explore the modelling of these different elements and their interplay.

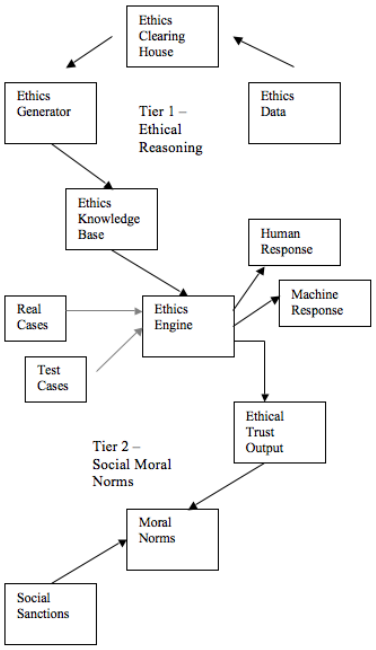

Figure 2: CMET 2-Tier Architecture

The CMET architecture is shown in Figure 2. This model corresponds to and is a conceptual expansion of the working mechanisms of the "change process" arrow marked in Figure 1. In the CMET model, tier-1 of the CMET architecture models ethical trust reasoning and tier-2 addresses the collective social moral norms.

3.3 Bio-inspired Neural Mechanisms

The internal mechanisms of CMET are drawn from two types of the bio-inspired neural systems. The first type is the generic neuron. The generic neuron is organized into artificial neural network (ANN) and deployed in tier-1 of CMET together with a multi-agent system (MAS) to account for social interactions. The second type is the mirror neuron system (MNS) and the concepts of MNS are deployed in MAS in tier-2 of CMET to account for required feedbacks and interactions.

Neurons are discrete cells found in the brain. Heinrich Wilhelm Gottfired von Waldeyer-Hartz first coined the term neuron in 1891 and later Cajal established the neuron doctrine and the principle of connection specificity. Cajal and Sherrington proposed that neurons contact at specialized points called synapses (See [13, 29, 45]). The neuron and synapse model was operationalised by [60]; for a history of artificial neural nets (ANN) see [51]. [39] improved on the ANN model by adding associative memory to the recurrent ANN. ANN have proved to be a suitable tool for classifications and pattern recognition. They will be used for classifications of ethical cases in tier-1 of CMET during ethical trust reasoning.

Mirror neuron is a new class of visuomotor neuron discovered in the premotor cortex of primates [25, 79]. Neural studies have identified the important role of mirror neurons systems (MNS) for action understanding and imitation [78, 79] and for normal social cognitive development [70, 82]. We envisage that this notion of MNS agents with the added "empathic role" [20, 24] of sensing and caring for other agents will be able to provide the necessary feedback link to allow collective moral actions to evolve from individual moral judgment and actions. According to [24], he reported two distinct classes of neurons in a particular sector of the premotor cortex: the canonical neuron and the mirror neuron. As part of [24] "shared manifold hypothesis", the canonical neuron 'simulates' the action required. The canonical neuron also 'simulates' the best "programmed plan" that will achieve the goal and provides a copy of the signal to the mirror neuron; this "simulation of the action is used to predict its consequences, thus enabling the achievement of a better control strategy" ([24] : p 40). This idea of simulation together with the theories of "mind-reading" [25] and other theories will be used in tier-2 of CMET.

Hence, we deploy two types of neural agents within our model as follows:

Tier-1 of CMET – Ethical Reasoning. In tier-1, we extend the use of 3-layer recurrent neural network (RNN) from the studies of [33, 34] for the classification of ethical cases; application of ethical rules and pragmatic schema and a multi agent-based system (MAS) for modelling social aspects of ethical trust reasoning. We envisage two possible forms of RNN and MAS linkages as shown in Figure 3 and Figure 4. In the loose coupling format (see Figure 3), the ANN is linked for shared data access. We expect to use this for rapid prototyping and concept testing.

The tight coupling linkage is use for investigation of the ethical trust reasoning once the concepts are tested and established. Here in the MAS, we have the following agents:

- Feature agent (F) corresponds to the input neurons of the traditional neural network. Feature agent holds the required feature information as well as the values of the synaptic weights. Feature agent can be a parent agent or when new features are required, a new child agent can be instantiated from existing parent feature agent. Hence, feature agent varies in number and unlike input neurons a feature agent allows the structure of the neural network to vary according to the ethical cases.

- Middle agent (M) carries out the function of the hidden layer of the neural network. A middle agent will apply the correct computational rule based on the numbers of feature agent present.

- Ethical agent (E) corresponds to the output neurons. This agent holds the outcomes of the ethical trust reasoning processes. Unlike the traditional output neurons, the pre-defined social parameters and other agents in the MAS can further influence the outcomes. This provide for social interactions and social influence on ethical trust reasoning.

- Logic agent (L) provides the input based on appropriate rules and logic system.

- Control agent (C) provides control and social influence to all the other agents.

The flow of this tight coupling is shown in Figure 4.

Tier 2 – Social Moral Norms. Here, MNS agents first transfer moral judgment into moral actions and next facilitate the

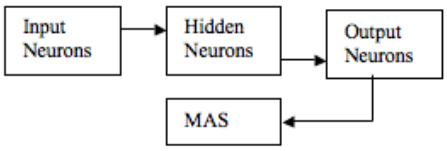

Figure 3: Loose Coupling of ANN with MAS

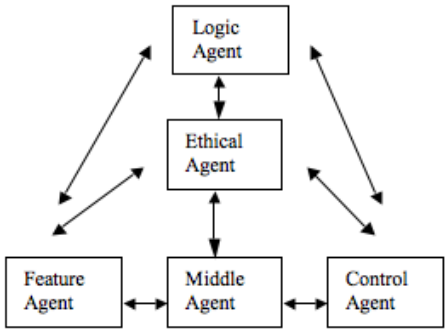

Figure 4: Tight Coupling of ANN within MAS

copying of the moral actions into collective moral practices. As shown in Figure 1, internal and external responses can provide feedbacks (shown as dotted blue arrows) to the internal and external stimuli respectively. To model this signal flow, we deploy appropriate agents with canonical and mirror functionality inspired by the MNS as reported in the previous sections. As seen in Figure 4, the control agent in the MAS implements the functions of canonical neurons; while the middle agent implement functions of the mirror neuron by applying the appropriate rules and adaptive learning that affect the formation of social moral norms. In addition, the implementation at this second tier draws from MNS and gesture communication theories [79]; psychological theories [85]; social cognitive theory of moral thought and action [10] and theories of adaptive structuration theories and multi-stage flow in communication studies. These theories inform the design of the collective moral norms of the CMET.

3.4 CMET Components

The key components of the CMET as shown in Figure 2 are as follows:

- Ethics Data encodes the ethical raw data into a format that can be used for computer and logic manipulations. The data is organised into appropriate data structures and stored in designated storage areas.
- Ethics Clearing House applies the various theories of normative ethics; ethics that are for arriving at standards and that regulate conduct. The various ethical datagrams are fed into the recurrent neural network (RNN) and grouped into appropriate ethical groups.
- Ethics Generator generates the ethical rule based on suitable forms of logic. Three types of logic are investigated including formal logic, fuzzy logic and abductive logic. The outputs from this component are appropriate ethical systems that form the ethics knowledge base.
- Ethics Knowledge Base contains the appropriate rules for each of the ethical systems. The knowledge from this component feeds the ethics engine. The knowledge base is updated for each of the ethics cases.
- Ethics Engine contains pragmatic schema and work together with the rules from the ethics knowledge base to extract appropriate "pragmatic" ethical outcomes from notions of upper and lower bound solution sets. This important step addresses and overcomes the traditional rationalist approach to reasoning that required a priori reasons be first present to have substantial understanding of the environment. The theory of Pragmatic Reasoning informs this area of design. The criteria adopted under the commonsense considerations include (a) Social acceptance; (b) Social convenience and (c) Social wisdom.
- Ethical Trust Output serves as the link to tier-2 of CMET. The outputs from real and test cases will be generated from the ethics engine and the responses in the form of human and machine responses are tuned to arrive at a suitable fit. These outputs are channelled to the moral norms module.
- Moral Norm simulates the required social network structures such as lattice, small world network and scale-free network. Mirror neural agents carry out appropriate functionality of "mirroring" and "adaptive learning" to generate the shape of the moral norms.
- Sanctions encapsulate the cost of sanctions and monitoring of moral norms that impact the shape of the moral norms. It serves to provide essential feedback to the system.

We implement components of CMET in a desktop computing environment using Java-based multi-agent simulation toolkits. These toolkits are compliant with the Foundation for Intelligent Physical Agent's specifications [22]. The two types of agent-based modeling and simulation (ABMS) toolkits deployed included (a) the Recursive Porous Agent Simulation Toolkit (Repast) (see [53]) and (b) Netlogo and Hubnet, the participatory simulation system [67]. Verification of the model and ensuring quality of the simulation model are important part of the design. For our project, we will deploy standard tools for checking codes for errors; use of suitable statistical techniques (t, Mann-Whitney) from simulation literatures for checking of output data; the interpretation of collective behavior that include "cluster analysis of agents" and the use of domain experts checking of simulated outputs – conceptually similar to that of a "Turing test" where we compare output data from the model to those from the system.

4. CONCLUSION

Neural mechanisms from the study of the human brain and neuroscience have excellent properties for classifications and cognition. In particular, these elements are useful for action understanding, caring and making sense of the actions of others based on their own internal system. These features

are useful for the study of ethical trust and social moral norms.

Specifically, we have adapted the computational features of neural nets in terms of artificial neural networks (ANN) for ethical trust reasoning and deployed them in tier-1 of CMET. We also implemented various functionalities of the mirror neuron systems (MNS) as feedback mechanisms of our MAS in tier-2 of CMET. The design of CMET is supported by key theories such as "shared manifold hypothesis" from mirror neuron systems; social signal theory; social cognitive theory of moral thought and action; and theories of adaptive learning. Our project is now at the rapid prototyping and concept-testing phase where we will gather data on appropriate ethical cases and populate the ethics knowledge base. Our next phase is a simulation of the ethical agents and a study of the emergent properties of social moral norms.

5. REFERENCES

[1] A. Abu-Akel. A neurobiological mapping of theory of mind. *Brain Research Reviews*, 43(1):29–40, 2003.

[2] R. Adolphs. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178, 2003.

[3] C. Allen, W. Wallach, and I. Smit. Why machine ethics? *Intelligent Systems, IEEE*, 21(4):12–17, 2006.

[4] M. Anderson and S. L. Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–25, 2007.

[5] M. Anderson, S. L. Anderson, and C. Armen. Towards machine ethics. In *Proceedings of AAAI Workshop on Agent Organizations: Theory and Practice, San Jose, CA, July*, 2004.

[6] M. Anderson, S. L. Anderson, and C. Armen. An approach to computing ethics. *Intelligent Systems, IEEE*, 21(4):56–63, 2006.

[7] S. W. Anderson, A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, 2:1032–1037, 1999.

[8] R. C. Arkin. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. GVU Technical Report GIT-GVU-07-11, 1-117, College of Computing, Georgia Tech, 2007.

[9] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.

[10] A. Bandura. *Social cognitive theory of moral thought and action*, chapter In W. M. Kurtines and J. L. Gewirtz (ed) Handbook of moral behavior and development: Theory, research and applications. Erlbaum, Hillsdale, NJ, 1991.

[11] J. Bolender. A two-tiered cognitive architecture for moral reasoning. *Biology and Philosophy*, 16(3):339–356, 2001.

[12] J. Borenstein. The ethics of autonomous military robots. *Studies in Ethics, Law, and Technology*, 2(1), 2008.

[13] C. S. Breathnach. Charles Scott Sherrington's integrative action: a centenary notice. *Journal of the Royal Society of Medicine*, 97:34–36, 2004.

[14] S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems, IEEE*, 21(4):38–44, 2006.

[15] C. J. Bullock. Trust literature: A survey and a criticism. *The Quarterly Journal of Economics*, 15(2):167–217, 1901.

[16] C. Castelfranchi and R. Falcone. *Social Trust: A Cognitive Approach, chapter In Cristiano Castelfranchi and Yao-Hua Tan (ed) Trust and Deception in Virtual Societies*. Kluwer Academic Publishers, Dordrecht/Boston/London, 2001.

[17] C. Castelfranchi, R. Falcone, and G. Pezzulo. Trust in information sources as a source for trust: a fuzzy approach. In *International Conference on Autonomous Agents, Proceedings of the second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia, SESSION: Social networks and trust*, 2003.

[18] P. Danielson. Competition among cooperators: Altruism and reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 2002.

[19] P. Danielson. Playing with ethics: Games, norms and moral freedom. *Topoi*, 24(2):221–227, 2005.

[20] F. de Vignemont and T. Singer. The empathic brain: how, when and why? *Trends in Cognitive Sciences*, 10(10):435–441, 2006.

[21] J. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity)*. Princeton University Press: Princeton, 2007.

[22] FIPA. Foundation for intelligent physical agent (FIPA) home page, http://www.fipa.org/. accessed 18th August 2008., 2008.

[23] F. Fukuyama. *Trust: The Social Virtue and the Creation of Prosperity*. Hamish Hamilton London, 1995.

[24] V. Gallese. The shared manifold hypothesis. from mirror neurons to empathy. *Journal of Consciousness Studies*, 8:33–50, 2001.

[25] V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12):493–501, 1998.

[26] D. Gambetta. *Can We Trust Trust*, chapter 13, pages 213–237. New York: Basil Blackwell, 2000.

[27] J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9(1):39–47, 2007.

[28] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers. The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*, 35(4):1674–1684, 2007.

[29] M. Glickstein. Golgi and Cajal: The neuron doctrine and the 100th anniversary of the 1906 nobel prize. *Current Biology*, 16(5):R147–R151, 2006.

[30] J. Greene and J. Haidt. How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12):517–523, 2002.

[31] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen. The neural bases of cognitive

conflict and control in moral judgment. *Neuron*, 44(2):389–400, 2004.

[32] N. Griffiths and M. Luck. Coalition formation through motivation and trust. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems. Melbourne, Australia.*, 2003.

[33] M. Guarini. Mind, morals, and reasons. *Lecture Notes in Computer Science, Practical Reasoning*, 1085:305–317, 1996.

[34] M. Guarini. Particularism and the classification and reclassification of moral cases. *Intelligent Systems, IEEE*, 21(4):22–28, 2006.

[35] T. W. Guinnane. Trust: A concept too many. Centre Discussion Paper 907, Economic Growth Center, Yale University, 2005.

[36] J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(1):814–834, 2001.

[37] J. Haidt. The new synthesis in moral psychology. *Science*, 316(5827):998–1002, 2007.

[38] M. Hauser, F. Cushman, L. Young, K.-X. Jin, and J. Mikhail. A dissociation between moral judgments and justifications. *Mind & Language*, 22(1):1–21, 2007.

[39] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982.

[40] L. T. Hosmer. Trust: The connecting link between organizational theory and philosophical ethics. *The Academy of Management Review*, 20(2):379–403, 1995.

[41] D. Johnson and K. Grayson. Cognitive and affective trust in service relationships. *Journal of Business Research*, 58(4):500–507, 2005.

[42] A. J. I. Jones. On the concept of trust. *Decision Support Systems*, 33(3):225–232, 2002.

[43] C. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. *Multi-Agent System Engineering*, pages 221–231, 1999.

[44] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[45] E. R. Kandel and L. R. Squire. Neuroscience: Breaking down scientific barriers to the study of brain and mind. *Science*, 290(5494):1113–1120, 2000.

[46] S. Knowles. Is social capital part of the institutions continuum and is it a deep determinant of development? Research Paper 2006/25, United Nations University - World Institute for Development Economics Research, 2006.

[47] M. Koenigs, L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, and A. Damasio. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138):908–911, 2007.

[48] J. E. LeDoux. Emotion: Clues from the brain. *Annual Review of Psychology*, 46:209–235, 1995.

[49] J. D. Lewis and A. Weigert. Trust as a social reality. *Social Forces*, 63(4):967–985, 1985.

[50] M. D. Lieberman. Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58(1):259–289, 2007.

[51] R. Lippmann. An introduction to computing with neural nets. *ASSP Magazine, IEEE* [*see also IEEE Signal Processing Magazine*], 4(2):4–22, 1987.

[52] N. Luhmann. *Trust and Power*. John Wiley and sons, 1980.

[53] C. M. Macal and M. J. North. Agent-based modeling and simulation: Desktop ABMS. In *Proceedings of the Winter Simulation Conference, Washington, D.C. USA*, 2007.

[54] T. Malsch and I. Schulz-Schaeffer. Socionics: Sociological concepts for social systems of artificial (and human) agents. *Journal of Artificial Societies and Social Simulation*, 10(1), 2007.

[55] S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Mathematics and Computer Science, University of Stirling, 1994.

[56] P. Massa. *A Survey of Trust Use and Modeling in Real Online Systems*, chapter 3, In Ronggong Song and Larry Korba and George Yee (ed) Trust in E-services: Technologies, Practices and Challenges. Idea Group, Inc., 2007.

[57] E. M. Maximilien and M. P. Singh. Agent-based trust model involving multiple qualities. In *International Conference on Autonomous Agents, Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems table of contents. SESSION: Papers: trust and reputation, Pages: 519 - 526*, 2005.

[58] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[59] D. J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, 38(1):24–59, 1995.

[60] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[61] D. H. McKnight and N. Chervany. Trust and distrust definitions: One bite at a time. *Trust in Cyber-societies*, pages 27–54, 2001.

[62] D. H. Mcknight, C. J. Kacmar, and V. Choudhury. Shifting factors and the ineffectiveness of third party assurance seals: A two-stage model of initial trust in a web business. *Electronic Markets*, 14(3):252–266, 2004.

[63] B. M. McLaren. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *Intelligent Systems, IEEE*, 21(4):29–37, 2006.

[64] J. P. Mitchell, M. R. Banaji, and C. N. Macrae. The link between social cognition and self-referential thought in the medial prefrontal cortex. *The Journal of Cognitive Neuroscience*, 17(8):1306–1315, 2005.

[65] G. Möllering. Understanding trust from the perspective of sociological neoinstitutionalism: The interplay of institutions and agency. Technical report, MAX PLANCK Institute for the study of Societies, MPIfG Discussion Paper 05/13, 2005.

[66] J. H. Moor. The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, 21(4):18–21, 2006.

[67] NetLogo. Netlogo home page,

http://ccl.northwestern.edu/netlogo/. accessed 18th August 2008, 2008.

[68] B. Nooteboom. *Trust: Forms, Foundations, Functions, Failures and Figures*. Edward Elgar, Cheltenham, UK, 2002.

[69] B. Nooteboom, T. Klos, and R. Jorna. Adaptive trust and co-operation: An agent-based simulation approach. *Trust in Cyber-societies : Integrating the Human and Artificial Perspectives*, pages 83–, 2001.

[70] L. M. Oberman and V. S. Ramachandran. The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin*, 133(2):310–327, 2007.

[71] E. Oztop, M. Kawato, and M. Arbib. Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19(3):254–271, 2006.

[72] G. Paperin, D. Green, S. Sadedin, and T. Leishman. A dual phase evolution model of adaptive radiation in landscapes. *Progress in Artificial Life*, pages 131–143, 2007.

[73] R. Parry. Ancient ethical theory. The Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/ethics-ancient/, accessed on 03 July 2008, 2004.

[74] L. Pereira and A. Saptawijaya. Modelling morality with prospective logic. *Progress in Artificial Intelligence*, pages 99–111, 2007.

[75] T. M. Powers. Prospects for a kantian machine. *Intelligent Systems, IEEE*, 21(4):46–51, 2006.

[76] K. Prehn, I. Wartenburger, K. Meriau, C. Scheibe, O. R. Goodenough, A. Villringer, E. van der Meer, and H. R. Heekeren. Individual differences in moral judgment competence influence neural correlates of socio-normative judgments. *Social Cognitive and Affective Neuroscience*, 3(1):33–46, 2008.

[77] A. Raine and Y. Yang. Neural foundations to moral reasoning and antisocial behavior. *Social Cognitive and Affective Neuroscience*, 1(3):203–213, 2006.

[78] G. Rizzolatti. The mirror neuron system and its function in humans. *Anatomy and Embryology*, 210:419–421, 2005.

[79] G. Rizzolatti and L. Craighero. The mirror neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

[80] J. Rotter. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4):651–665, 1967.

[81] F. D. Schoorman, R. C. Mayer, and J. H. Davis. An integrative model of organizational trust: Past, present, and future. *The Academy of Management Review (AMR)*, 32(2):344–354, 2007.

[82] M. Siegal and R. Varley. Neural systems involved in theory of mind. *Neural Reviews Neuroscience*, 3:463–471, 2002.

[83] G. Simmel. *The Sociology of Georg Simmel - Translated and Edited by Kurt H Wolff*. Free Press, New York, 1950.

[84] A. Smith. *The Theory of Moral Sentiments, [Adam Smith, 1723-1790], edited by D. D. Raphael and A. L. Macfie*, volume 1 of *LibertyClassics*. LibertyClassics, 1982.

[85] E. Thompson. Empathy and consciousness. *Journal of Consciousness Studies*, 8:1–32, 2001.

[86] L. Q. Uddin, M. Iacoboni, C. Lange, and J. P. Keenan. The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4):153–157, 2007.

[87] J. van den Hoven and G.-J. Lokhorst. Deontic logic and computer-supported computer ethics. *Metaphilosophy*, 33(3):376–386, 2002.

[88] G. Vlastos. *Socrates, Ironist and Moral Philosopher*. Cornell University Press, Ithaca, NY, 1991.

[89] W. Wallach. Implementing moral decision making faculties in computers and robots. *AI & Society*, 22(4):463–475, 2008.

[90] K. Warwick. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology*, 5(3):131–137, 2003.

[91] R. Waterfield. *The First Philosophers: The PreSocratics and Sophists*. Oxford University Press, New York, USA, 2000.

[92] A. C. Wicks, S. L. Berman, and T. M. Jones. The structure of optimal trust: Moral and strategic implications. *The Academy of Management Review*, 24(1):99–116, 1999.

[93] O. E. Williamson. Calculativeness, trust, and economic organization. *Journal of Law and Economics*, 36(1):453–486, 1993.