A Review of Computational Models of Trust

Stephen Cranefield
University of Otago
stephen.cranefield@otago.ac.nz

Steve Reeves
University of Waikato
stever@waikato.ac.nz

September 2023

1 Introduction

Trust is a key component of human business and social transactions, and has been widely studied in the social sciences. This article presents a brief overview of different definitions of trust, before focusing in the next section on computational models of trust. For a discussion of other accounts of trust from a social science perspective see Marsh [14, §3.2].

2 Definitions of trust

We present a few of the "significant definitions of trust" that Castelfranchi and Falcone critique [4]. We include additional definitions, dimensions and characteristics of trust identified in surveys on trust modelling by Ramchurn et al. [19], Jøsang et al. [12], and Cho et al. [5], as well as the view of Castelfranchi and Falcone.

- **Trust as a predictability**: "Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends" [9, translated from Italian]. Jøsang et al. refer to this as *reliability trust* and note that it combines the concepts of *dependence* on the trusted party as well as an assessment of their *reliability*.
- **Trust as the decision to make oneself vulnerable**: "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"[15].

  McKnight and Chervany [16] present a similar definition of *trusting intention*: "the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible". They state that this this intention depends on *trusting beliefs*, especially about the other party's benevolence, honesty, competence, and predictability.

  Jøsang et al. consider that this *decision trust* adds to the previous definition's notions of dependence and reliability the additional considerations of the *utility* of positive and negative outcomes of the trust decision and the *risk attitude* of the trusting party.
- **Trust as expectation**: x trusts y if and only if "x expects that y will behave according to x's best interest, and will not attempt to harm x" [14, p.56].
- **Trust based on reciprocity**: "the willingness to take some risk in relation to other individuals on the expectation that the others will reciprocate" [17].
- **Trust as encapsulated interest**: "I trust you because I think it is in your interest to take my interests in the relevant matter seriously" [11].

- **Trust as a composite psychological state**: "[Trust is] a psychological state of a trustor comprising the intention to accept vulnerability in a situation involving risk, based on positive expectations of the intentions or behavior of the trustee" [20].
- **Socio-cognitive trust**: Castelfranchi and Falcone present their own view of trust as follows:

  > *. . . trust is a* relational *construct, involving at the same time:*
  >
  > – *A subject X (the trustor) which necessarily is an 'intentional entity', i.e. a system that we interpret according to Dennett's intentional stance [7], and that is thus considered a cognitive agent.*
  > – *An addressee Y (the trustee) that is an agent in the broader sense of this term [3], i.e. an entity capable of causing some effect as the outcome of its behavior.*
  > – *The causal process itself (the act, or performance) and its result; that is, an act α of Y possibly producing the desired outcome O.*
  >
  > *Moreover, . . . trust is a layered notion, used to refer to several different (although interrelated) meanings . . . :*
  >
  > – *in its basic sense, trust is just a mental and affective attitude or disposition towards Y, involving two basic types of beliefs: evaluations and expectations;*
  > – *in its richer use, trust is a decision and intention based on that disposition;*
  > – *as well as the act of relying upon Y's expected behavior;*
  > – *and the consequent social relation established between X and Y.*
  >
  > (CastelFranchi and Falcone [4])

  In short, they state that "The basic nucleus of trust—as a mental disposition towards Y—is a positive expectation based on a positive evaluation; plus the idea that X might need Y's action".

  We note that the establishment (and maintentance) of a social relation, mentioned above, is an important factor in our context of interest: interactions in value chains, and is not explicitly highlighted in many definitions of trust.

3 Trust decisions

Cho et al. [5, Fig. 1] present an abstract view of a *trust assessment process* in which an entity i considers whether to trust another entity j to perform a task. This assessment is influenced by *individual* and *relational* trust attributes. Individual trust attribues include *logical* factors such as beliefs, confidence, experience and rationality, which inform asessments of trust based on evidence and logical or mathematical reasoning, and *emotional* factors such as expectation, willingness and propensity. Relational trust attributes include measures of *similarity* between the trustor and the candidate trustee, *social metrics* such as the candidates's centrality and betweenness, and the *importance* of that party, e.g. as measured by the Shapley value from cooperative game theory. The assessment process then considers risk factors such as information uncertainty, the trustor's vulnerability to failure and the impact of a failure, before calculating the potential gain and loss that could arise from the trust decision and choosing to trust if the gain outweighs the loss[1]. Finally, the observed outcome is used to adjust the "prior belief" in the candidate trustee's trustworthiness.

Jøsang et al. discuss four "trust classes" (actually specific situations in which trust decisions are made) identified by Grandison and Sloman [10] in research on trust in internet applications. The two presentations do not appear to match completely, so we combine the two lists to give five trust decision points, where the third and fourth appear in only one of the cited works. For the first four classes, we use the terminology of Jøsang et al.

[1] Risk is not explicitly included in the description of this decision. We assume the intention is to compare the *expected* gain and loss.

- *Access trust*: trusting another party to access the trustor's resources.
- *Provision trust*: trusting a service or resource provider.
- *Delegation trust*: trusting an agent to make decisions for the trustor.
- *Identity trust*: trusting the asserted identity of the trustee (Jøsang et al.).
- *Certification trust*: trusting a trustee based on its certification by a third party (Grandison and Sloman).
- *System trust*: trusting that "the necessary systems and institutions are in place in order to support the transaction and provide a safety net in case something should go wrong" [12].

Jøsang et al. refer to system trust as *context trust* and Grandison and Sloman call this *infrastructure trust*. However, we use the terminology of McKnight and Chervany [16], who define it as trust in "the presence of structural assurances based on regulations, guarantees, contracts, etc. and the perception that the system is in a normal operational state".

Cho et al. identify four sources of information that can inform trust decisions in computational systems, which they call *dimensions* of trust:

> - *Communication trust from communication networks such as quality of service (e.g., service response time, packet drop rates).*
> - *Information trust from information networks (e.g., information credibility, veracity, integrity).*
> - *Social trust from interactions/networks (e.g., a source's reliability)*
> - *Cognitive trust from cognitive process (e.g., cognitive information processing capability)*
>
> (Cho et al.[5])

4 Computational models of trust

A wide range of computational models for assessing trust and/or reputation have been presented in the literature. In this section we present some examples of few types of model in order to give the flavour of this field of research. A wider coverage can be found in the systematic literature reviewss by Pinyol and Sabater-Mir [18] and Braga et al. [2], which classify published computational trust models across various dimensions.

The study of trust in computational systems was pioneered by Marsh [14]. After reviewing social science literature on trust, Marsh proposed a mathematical model of the trust an agent x should have of agent y in situation α:

$$T_x(y, \alpha) = U_x(\alpha)\ I_x(\alpha)\ \widehat{T_x(y)}$$

where $U_x(\alpha)$ and $I_x(\alpha)$ are, respectively, the utility and importance of situation α to agent x, and $\widehat{T_x(y)}$ is an estimate of the situation-independent trust that x has of agent y by aggregating historic $T_x(y, \alpha)$ values.

Marsh discusses different different trusting dispositions (optimistic, pessimistic and realistic) that could affect the calculation of $\widehat{T_x(y)}$, proposes a model for how agents can choose to cooperate with other agents (based on notions of risk and competence), and discusses how trust estimates could be modified in reciprocal relationships based on interaction history. The use of this model is illustrated in a furniture moving scenario in which two agents must cooperate with each other to complete their assigned furniture moving tasks. Simulation experiments were also performed to study how the performance of agents with different trusting dispositions performed as expected, and to determine the conditions under which trusting behaviour can be learned.

Ramchurn et al. [19] review research on trust in the context of multi-agent systems: open distributed systems of autonomous agents that "act and interact in flexible ways to achieve their design objectives in uncertain and dynamic environments". They adopt the following definition of trust, adapted from Dasgupta [6]:

Trust is a belief an agent has that the other party will do what it says it will (being honest and reliable) or reciprocate (being reciprocative for the common good of both), given an opportunity to defect to get higher payoffs.

The review considers approaches for:

- *individual-level* trust, where either a) trust emerges as agents individually learn cooperative interaction strategies to maximise their payoff, or b) agents individually maintain trust and reputation models of other agents to inform their interaction decisions; and
- *system-level* trust, where the multi-agent system as a whole is designed to protect against malicious behaviour. This can be subdivided into three catagories: a) adopting the use of communication protocols that incentivize truth-telling (e.g. the use of Vickrey-Clarke-Groves auctions), b) providing system-level reputation-tracking services, and c) adopting security mechanisms that ensure new entrants to the multi-agent system can be trusted.

We do not comment further on 1a, 2a and 2c, as these seem less applicable to the value chain context than approaches 1 and 2b.

Under individual-level trust, Ramchurn et al. discuss research on *trust metrics* for quantifying the trustworthiness of other agents based on past interactions, *reputation models* that allow agents to aggregate ratings of potential partners from their own experience and by other agents in their social network, and *socio-cognitive* models of trust, in which trust decisions are informed not by quantitative models of other agents but by explicit cognitive models that include beliefs about other agents, such as their competence and motivation. We discuss these in more detail below, along with use of *reputation mechanisms* at the system level.

4.1 Trust metrics

Trust metrics may be specific to a particular problem domain or generic. An an example of a problem-specific metric, Ramchurn et al. discuss the work of Witkowski et al. [22], who consider agents that trade bandwidth in an intelligent telecommunications network. In this scenario, trust in other agents is a function of the difference between agents' bids and the received bandwidth in past trades (for consumers) and the degree to which the provided bandwith was exploited (for suppliers). In contrast, general trust metrics that are based on assessments of agents as either good or bad are more general but are criticised by Ramchurn et al. for lacking the richness of outcome assessment needed in realistic settings. As an example of a richer but more generally applicable approach, Ramchurn et al. discuss the REGRET model [21]. This is a reputation model (described in more detail below) that considers assessments of an agent's performance from multiple agents. Its generality is due to an ontological dimension that helps agents to understand each other's rating when they have different preferences (e.g. low price vs. high quality).

4.2 Reputation models

Ramchurn et al. describe three complementary aspects of reputation systems: the gathering of ratings about the trustworthiness of other agents using social relationships, the aggregation of these ratings, and mechanisms to promote accurate ratings. To give the flavour of models for aggregating ratings, we present some technical details of REGRET as presented by Sabater and Sierra [21], with some minor changes of notatation for brevity.

REGRET amalgamates *impressions* from multiple agents regarding the *outcome* of a dialogue between agents, which consists of an initial contract together with the actual result. An impression is modelled as a tuple $\iota = (a, b, o, \varphi, t, W)$ where agent a is providing a subjective opinion $W \in [-1, 1]$ of an interaction with agent b that had outcome o. The impression relates to a specific variable φ of the outcome, and t is the time that the impression is recorded.

Suppose $I_{a,b}(\varphi)$ is a set of impressions that agent a has recorded about agent b for outcome variable φ over a period of time. The *individual* reputation that a has of b at time t is modelled as a weighted average

of these impressions:

$$R_{a\to b}(\varphi) = \sum_{\iota_i \in I_{a,b}(\varphi)} \frac{f(t_i, t)}{\sum_{\iota_j \in I f(t_j,t)}} W_i$$

where $f(t_i, t)$ is a function that produces a higher value the closer that t_i is to t.

The *reliability* $RL_{a\to b}(\varphi)$ of this reputation is then calculated as a weighted sum of the number of impressions used to calculate it (denoted $Ni_{a\to b}(\varphi)$) and the variability of these impressions ($Dt_{a\to b}(\varphi)$):[2]

$$RL_{a\to b}(\varphi) = (1-\mu)\, Ni_{a\to b}(\varphi) + \mu\, Dt_{a\to b}(\varphi)$$

REGRET also considers three *social measures* of reputation and their associated reliabilities:

$R_{a\to\mathcal{B}}(\varphi)$**:**

A group $\mathcal{B}$'s reliability as assessed by an agent a, based on interactions between a and members of $\mathcal{B}$.

$R_{\mathcal{A}\to b}(\varphi)$**:**

The reputation of an agent b by a group $\mathcal{A}$, based on an aggregating the reputations of b held by members of $\mathcal{A}$.

$R_{\mathcal{A}\to\mathcal{B}}(\varphi)$**:**

The reputation of a group $\mathcal{B}$ from the viewpoint of a group $\mathcal{A}$, based on aggregating the individual reputations of members of $\mathcal{B}$ by members of $\mathcal{A}$.

REGRET then defines the overall *social reputation* of b to a as a weighted sum of the individual reputation $R_{a\to b}(\varphi)$ and the three social measures above, along with a corresponding measure of the reliability of this reputation.

Finally, the model is generalised to consider the use of multiple interaction outcome variables , given an ontology that models dependencies between different these variables.

Reputation models have also been proposed in the context of peer-to-peer networks, for example, Eigentrust [13] (not discussed by Ramchurn et al.) uses a metric for the *local* trust that peer a has in peer b by subtracting the number of unsatisfactory transactions that a has had with b from the number of satisfactory transactions. This is then normalised across all peers that a has interacted with. The Eigentrust algorithm is described succinctly by Afanafor et al. [1]:

> *Eigentrust assumes that trust is transitive (if A trusts B and B trusts C then A should trust C), and that an agent's ability to provide reputational information is correlated to its ability to perform a task. It then represents the (direct) trust relationship between agents in a matrix. Multiplying a vector by this matrix yields a new vector representing direct trust, and—under the assumptions described above—further multiplications of the resultant vector by the matrix yield a new vector describing first hand, second hand, and so on reputational information.*

Afanafor et al. extend this model to take into account the difference between the lack of information about an agent and the lack of trust in that agent.

4.3 System-level reputation mechanisms

The reputation models discussed above allow an agent to collect ratings of other agents by themselves and other agents and use this to inform future trust decisions. However, Ramchurn et al. [19] note that self-interested agents may have no motivation to share information, and that research on localised agent reputation models does not consider how the existence of reputation models could be used to incentivize trustworthy behaviour amongst agents in a society. This can be achieved by providing *reputation mechanisms* at the system level. These involve centralised or distributed entities that store and publish ratings of agents by their interaction partners, and protocols for using them that are designed to make it rational for agents to provide truthful ratings. Ramchurn et al. discuss several such mechanisms proposed in the literature.

[2] A precise definition of the variability is not provided by Sabater and Sierra [21].

Potential Goal	$\mathbf{G_0} : \mathsf{Goal}_X(g) = g_X$ with $g_X \subseteq p$	
Potential Expectation	$\mathbf{B_1} : \mathsf{Bel}_X(\mathsf{Can}_Y(\alpha, p))$ $\mathbf{G_1} : \mathsf{Will}_X(\mathsf{Can}_Y(\alpha, p))$	**(Competence)**
Potential Expectation	$\mathbf{B_2} : \mathsf{Bel}_X(\langle \mathsf{WillDo}_Y(\alpha) \rangle p)$ $\mathbf{G_2} : \mathsf{Will}_X(\langle \mathsf{WillDo}_Y(\alpha) \rangle p))$	**(Disposition)**
$\mathbf{B_3} : \mathsf{Bel}_X\ \mathsf{Dependence}_{XY}(\alpha, p)$		**(Dependence)**
$\mathbf{I} : \mathsf{Intend}_X(\mathsf{Rely\text{-}upon}_{XY})\ \tau$ (where τ is the execution of α with result g)		
$\mathbf{I_1} : \mathsf{Intend\text{-}that}_X(\langle \mathsf{Achieve}_Y(\alpha) \rangle p)$		
$\mathbf{I_2} : \mathsf{Intend}_X(\neg \langle \mathsf{Achieve}_{X \text{ or } Z}(\alpha) \rangle p)$ (where $Z \neq Y$)		
$\mathbf{I_3} : \mathsf{Intend}_X(\neg \mathsf{Do}_X(\alpha'))$ (where α' is some action interfering with α)		
$\mathbf{I_4} : \mathsf{Intend}_X(\mathsf{Do}_X(\alpha'') \rightarrow \mathsf{Do}_Y(\alpha))$ (where α'' is an explicit request to Y to do α)		

Figure 1: Mental ingredients for strong delegation (Castelfranchi and Falcone [4])

4.4 Socio-cognitive models of trust

Socio-cognitive models of trust are based on the conviction that trust is necessarily dependent on an agent's beliefs and goals. These models aim to provide a logical theory explaining how explicit reasoning about mental attitudes (such as goals, intentions and beliefs such as those above) can be used to decide when to trust another.

Falcone and Castelfranchi [8] assert that because trust is relative to an agent's goals and desires, only a cognitive agent can really 'trust' another, and that trust is itself a complex mental attitude that consists of beliefs. In their book on trust theory, Castelfranchi and Falcone [4] provide an analysis and pre-formal model of trust based on these ideas. They consider three types of belief that the trustor X holds about the trustee Y, characterised as follows:

1. "X believes that Y is able and well disposed (willing) to do the needed action."
2. "X believes that in fact Y will appropriately do the action, as she wishes."
3. "X believes that Y is not dangerous; therefore she will be safe in the relation with Y, and can make herself less defended and more vulnerable."

The first and third beliefs above record *evaluations* of Y by X, and the second and third beliefs record *expectations* about Y's behaviour relative to X's goal.

In more detail, Figure 1 presents an analysis from Castelfranchi and Falcone [4] of the components required for "strong delegation" of an action α by agent X to agent Y in order to achieve a state of the world p that includes a goal g_X of agent X. Strong delegation is defined as delegation that involves explicit agreement between the two parties. Even though we do not expect the reader to understand the notation used, it should be apparent that a decision to delegate in this approach involves X reasoning about a combination of mental attitudes:

- a specific goal g_X,
- beliefs about the dependence of X on Y's performance of α,
- intentions (goals that the agent is internally committed to) that establish and support a relience on Y, and
- beliefs about Y's competence and disposition in relation to performing α.

In their book, Castelfranchi and Falcone elaborate on the reasoning required to produce trust using mental attitudes such as these.

References

[1] Juan Afanador, Nir Oren, Murilo S. Baptista, and Maria Araujo. "From Eigentrust to a Trust-Measuring Algorithm in the Max-Plus Algebra". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence*. Vol. 325. Frontiers in Artificial Intelligence and Applications. IOS Press, 2020, pp. 3–10. DOI: 10.3233/FAIA200069.

[2] Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto. "Survey on Computational Trust and Reputation Models". In: *ACM Computing Surveys*, 51(5), 2019, pp. 1–40. DOI: 10.1145/3236008.

[3] C. Castelfranchi. "Towards an agent ontology: autonomy, delegation, adaptivity". In: *AI*IA Notizie*, 11(3), 1998, pp. 45–50.

[4] Cristiano Castelfranchi and Rino Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley & Sons, 2010.

[5] Jin-Hee Cho, Kevin Chan, and Sibel Adali. "A Survey on Trust Modeling". In: *ACM Computing Surveys*, 48(2), 2015, pp. 1–40. DOI: 10.1145/2815595.

[6] Partha Dasgupta. "Trust as a Commodity". In: *Trust: Making and Breaking Cooperative Relations*. Ed. by Diego Gambetta. Blackwell, 1988, pp. 49–72.

[7] Daniel C. Dennett. *The Intentional Stance*. Cambridge, MA: The MIT Press, 1987.

[8] Rino Falcone and Cristiano Castelfranchi. "Social Trust: A Cognitive Approach". In: *Trust and Deception in Virtual Societies*. Ed. by Cristiano Castelfranchi and Yao-Hua Tan. Springer Netherlands, 2001, pp. 55–90. DOI: 10.1007/978-94-017-3614-5_3.

[9] Diego Gambetta. "Can We Trust Trust?" In: *Trust: Making and Breaking Cooperative Relations*. Ed. by Diego Gambetta. Blackwell, 1988, pp. 213–237.

[10] Tyrone Grandison and Morri Sloman. "A Survey of Trust in Internet Applications". In: *IEEE Communications Surveys & Tutorials*, 3(4), 2000, pp. 2–16. DOI: 10.1109/COMST.2000.5340804.

[11] Russell Hardin. *Trust and Trustworthiness*. The Russell Sage Foundation Series on Trust 4. Russell Sage Foundation, 2002.

[12] Audun Jøsang, Roslan Ismail, and Colin Boyd. "A Survey of Trust and Reputation Systems for Online Service Provision". In: *Decision Support Systems*, 43(2), 2007, pp. 618–644. DOI: 10.1016/j.dss.2005.05.019.

[13] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. "The Eigentrust Algorithm for Reputation Management in P2P Networks". In: *Proceedings of the Twelfth International Conference on World Wide Web - WWW '03*. ACM Press, 2003, p. 640. ISBN: 978-1-58113-680-7. DOI: 10.1145/775152.775242.

[14] Stephen Paul Marsh. "Formalising Trust as a Computational Concept". PhD thesis. University of Stirling, 1994. URL: http://stephenmarsh.wdfiles.com/local--files/start/TrustThesis.pdf.

[15] Roger C. Mayer, James H. Davis, and F. David Schoorman. "An Integrative Model Of Organizational Trust". In: *Academy of Management Review*, 20(3), 1995, pp. 709–734. DOI: 10.5465/amr.1995.9508080335.

[16] D.H. McKnight and N.L. Chervany. *The Meanings of Trust*. Technical Report 96-04. University of Minnesota, Management Information Systems Reseach Center, 1996.

[17] Elinor Ostrom and James Walker, eds. *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. Russell Sage Foundation Series on Trust v.6. Russell Sage Foundation, 2003. ISBN: 978-0-87154-647-0.

[18] Isaac Pinyol and Jordi Sabater-Mir. "Computational Trust and Reputation Models for Open Multi-Agent Systems: A Review". In: *Artificial Intelligence Review*, 40(1), 2013, pp. 1–25. DOI: 10.1007/s10462-011-9277-z.

[19] Sarvapali D. Ramchurn, Dong Huynh, and Nicholas R. Jennings. “Trust in Multi-Agent Systems”. In: *The Knowledge Engineering Review*, 19(1), 2004, pp. 1–25. DOI: 10.1017/S0269888904000116.

[20] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. “Not So Different After All: A Cross-Discipline View Of Trust”. In: *Academy of Management Review*, 23(3), 1998, pp. 393–404. DOI: 10.5465/amr.1998.926617.

[21] Jordi Sabater and Carles Sierra. “REGRET: Reputation in Gregarious Societies”. In: *Proceedings of the Fifth International Conference on Autonomous Agents - AGENTS ’01*. ACM Press, 2001, pp. 194–195. ISBN: 978-1-58113-326-4. DOI: 10.1145/375735.376110.

[22] Mark Witkowski, Alexander Artikis, and Jeremy Pitt. “Experiments in Building Experiential Trust in a Society of Objective-Trust Based Agents”. In: *Trust in Cyber-Societies*. Ed. by Rino Falcone, Munindar Singh, and Yao-Hua Tan. Vol. 2246. Springer Berlin Heidelberg, 2001, pp. 111–132. DOI: 10.1007/3-540-45547-7_7.