A Survey on Trust in Autonomous Systems

Shervin Shahrdar and Mehrdad Nojoumian

Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University, Boca Raton, FL, USA
sshahrda@fau.edu, www.shervinshahrdar.com
mnojoumian@fau.edu, http://faculty.eng.fau.edu/nojoumian/

Abstract. As a result of the exponential growth in technology and computing in the past couple of decades, autonomous systems are becoming more relevant in our daily lives. As these autonomous systems evolve and become more complex, the concept of trust in such systems becomes a major challenge that affects the performance, and reliability of such systems. Many prior studies have indicated that currently, humans have a very low trust-level in the fully autonomous robots. Similarly, trust between autonomous systems plays a major role in their performance. In this meta-analysis, we will explore various research and trust models in order to show why trust management is a very challenging aspect of future AI technologies.

Keywords: Trust Function, Reputation Systems, Autonomous Systems, Multi-Agent Systems

1 Introduction

The rapid growth in technology has resulted in the automation of many day to day tasks that humans had to perform themselves just decades ago. From ATMs (Automated Telling machines) to industrial robots used in factories, automation continues to aid humans with repetitive, difficult, and monotonous tasks. This technological advancement introduces newer and more complex robotic concepts, and automated systems in different areas of our lives including our homes, daily commute, workspace, military, and many other areas every day.

As these automated systems evolve, their levels of complexity increase overtime, and their involvement in various aspects of human life introduces the concept of human-robot trust. Studies have indicated that one of the most important challenges for successful integration of advanced autonomous systems and AI technology in human civilization, will be the management and the development of this mutual trust [8].

In this paper, we will provide a step by step, comprehensive analysis and explore various concepts and related studies, as well as experimental techniques proposed by researchers in this field, in order to understand how the human-robot trust management works, and how we can improve it. Additionally, we will explore the trust between autonomous systems (Agents) in section 3.

2 Definitions

2.1 What is an Autonomous System?

First, an autonomous system needs to be defined. This definition keeps changing everyday as the related technology grows exponentially [40].

Merriam Webster dictionary defines 'autonomy' as "The quality or state of being self-governing; especially: the right of self-government". The concept of Autonomy has existed for thousands of years, in many different areas including philosophy, sociology, politics, and technology. In fact, the second part of the term Autonomous, *nomos*, means 'law' in Greek. An autonomous entity "creates its own laws". [5].

We need to narrow down this definition, as this it is very broad and applies to many different areas and concepts. Perhaps the term that we are looking for is 'autonomous robots'. Autonomous robots are defined as, "intelligent machines capable of performing tasks in the world by themselves, without explicit human control." [9]

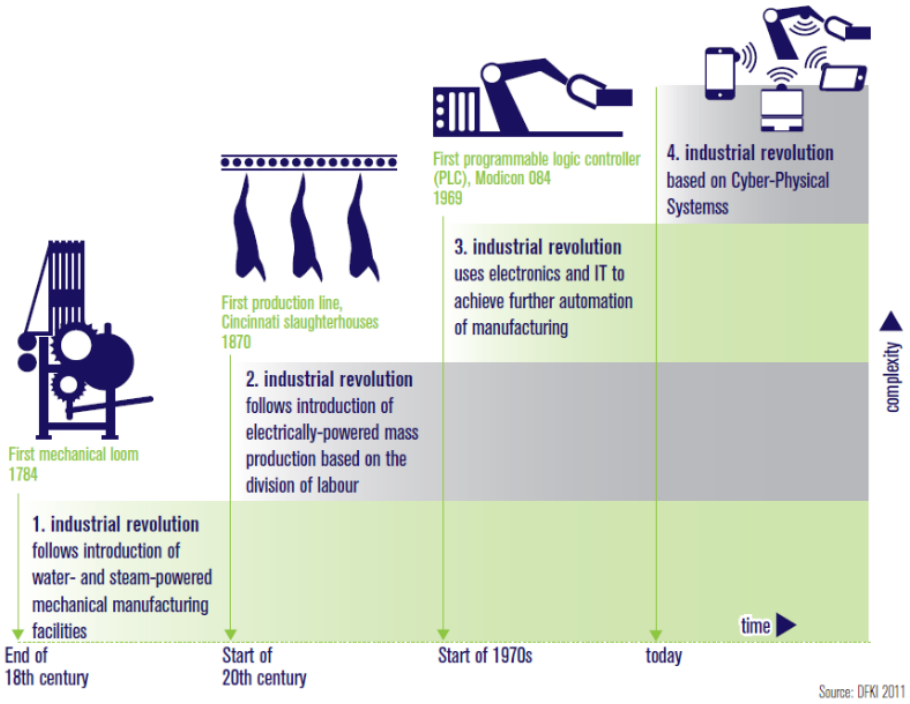

Fig. 1. Evolution of Automation [24]

Although automation was introduced to human civilization many years ago, and widely used after industrial revolution in early 1800s [39], Autonomous sys-

tems, and Industrial Robots are relatively new terms that were introduced merely decades ago, and their definitions are evolving everyday, as technology advances. In this paper, our main focus will be on Autonomous Robots and Autonomous Agents which are a highly advanced form of automated machines that have a high degree of self-awareness, and are capable of independently performing tasks which were previously done by humans.

2.2 What is Trust?

'trust' is a concept that has many different definitions in various contexts such as psychology, sociology, and Economics. Currently, there is no uniform definition of 'trust' in the context of psychology, or other areas [2]. Prior research indicates that there are over 300 definitions in various research areas, and in the context of Human-Robot relationships, which is our focus in this paper, there are more than 30 definitions. These definitions include 'human interpersonal trust', 'automation trust', 'trust in software agents", and many others [40].

Although there are complex definitions of trust documented based on the area of the research, we shall focus on the simplest, domain-agnostic definition. A very simple, generic definition of 'trust' would be: "A firm belief in the reliability, truth, or ability of someone or something" [44]. If we consider this simple definition, then, in our research context, the definition of trust would be: "A strong human belief in the reliability, truth, or ability of an autonomous system". The 'Autonomous system' in this case could be any kind of a self aware machine that has a high degree of autonomy. Some concrete examples would be human trust in self driving cars (SDC), autonomous planes, battlefield robots, rescue robots, autonomous software agents, and so on.

3 Review of Literature: Trust in Autonomous Systems

In this section, we will review, and categorize previous studies that are in the domain of human-robot trust relationship.

3.1 Trust Between Humans and Autonomous Systems

3.1.1 Human-Robots As previously stated, many studies have shown that currently, the level of trust between humans and autonomous systems is very low. That means in serious situations, humans tend to not let fully autonomous systems take control. A study that explores the low level of human-robot trust, is done by Daniel Stormont (2008) [42]. In this study, the factors that affect the trust between humans and robotic systems are analyzed. The study concluded that it is not just 'trust' that determines the usability, but 'confidence' also plays a significant role. One of the reasons the level of confidence in autonomous systems is very low is due to their low level of reliability. As cited in this paper, "A 2004 study of commercially available ruggedized robots operating in field conditions showed a mean-time-between-failures (MTBF) of 12.26 hours and an

availability rate of 37%"[12]. This means, that if the robotic systems reduce their failure rate, their reliability will increase, thus, the human confidence and trust in them will increase. Of course this study was done in 2004. It is clear that due to recent advancements in robotic systems, the MTBF should be definitely less than 37%. For example, when the first fatal car accident was reported for Tesla self driving cars, it became a topic of discussion in the media. Although these accidents are rare, companies like Tesla and Google learn the nature of the accident, and work on reducing the failure rate in the future models. [41]

In his research, Mr.Stormont (2008) also believes that another factor affecting the trust between humans and fully autonomous systems would be their unpredictability. It is known that in various hazardous circumstances such as battlefields and rescue missions, the unpredictability of robots becomes a critical problem for human supervisors. It has been argued that the autonomous nature of robots, and their decisions making could be a positive trait, since they could react faster to certain dangers compared to humans, but the problem arises when life and death of humans will depend on the choices of a robot; Questions such as "should life and death decisions be made by an autonomous system?" will emerge. In the 2008 study, a simulation of robots assisting firefighters in a hazardous fire situation was discussed. It was concluded that even though firefighters did not initially trust these helper robots, as the mission went on, and they got tired, their reliance and trust in the robots increased, as they helped them extinguish the fire.

Fig. 2. Example of autonomous and semi-autonomous robots used in military [4]

Another study conducted by Mr. Babak Esfandiari and his colleague, Mr. Sanjay Chandrasekharan [15] thoroughly examined and proposed "simple mechanisms for trust acquisition based on a very basic and general definition of trust." According this this study, the majority of views on the definition of trust are divided into two groups: Cognitive views, and mathematical views. Mr. Esfandiari argues that both of these views have something in common: both of them see 'trust' as a variable, that also has a threshold for an action. This action is called cooperation. The mathematical definition of 'trust' provided by this study

is: "Trust is a function T between any two agents of a set A of agents". These agents can be humans, or robots, or autonomous systems, and so one. This study also provides a concept called 'Trust Acquisition'. Given an example of previously defined mathematical function: T(Alice, Bob) (Trust between Alice and Bob), Trust Acquisition is described as "the process or mechanism that allows the calculation and update of T. In our definition, acquisition is not necessarily an 'increase' of T.". This study provides various methods of Trust Acquisition:

1. Trust Acquisition by Observation
   Mr. Esfandiari argues that Trust Acquisition can be obtained by performing 'Bayesian Learning'. This means that agents observe, and consider that past actions of other agents, and decide whether to perform an action or not. This paper provides an example of two agents, RoboCop robots John and Mary. John has the ball, and is deciding whether or not pass the ball to Mary, or just shoot the ball himself. If John is able to review the past performances of Mary, and compares them with his own performance, and performs a statistical analysis, he will be able to make a decision.
2. Trust Acquisition by Interaction
   In Trust Acquisition by Interactions protocol, agent 1 asks a bunch of predetermined questions from agent 2. Agent 1 already knows the answer, thus, the trust will be acquired based on the the number of correct answers provided by agent 2.
3. Trust Acquisition Using Institutions This study provides an example of humans trusting police officers wearing uniforms. If a person sees another person equipped with a gun who is wearing a uniform, he will automatically trust that person, because the trust is already established by the institution. However, if he sees a person not wearing a uniform and carrying a gun, he needs to make more calculations to trust this person.

In a 2011 paper, Peter A. Hancock et al. provided a comprehensive analysis of factors affecting trust in "Human-Robot Interaction (HRI)" [18]. This study categorizes factors that affect the trust in HRI into three different categories: human related, Robot related, and environmental variables, and each category has sub categories. Human related factors include training, expertise, situational awareness, demographics etc. Similarly, robot-related factors are behavior, dependability, reliability, level of automation, failure rates, false alarms, and transparency, and attribute based factors such as location, personality, adaptability, robot type, and Anthropomorphism (having human traits). Environmental factors include team work, culture, communication, shared mental models, task type, task complexity, multi-tasking, physical environment. This paper discovered that robot performance has the biggest impact on trust in the context of HRI, and tweaking the robot performance has a direct impact on trust. For example, if an autonomous robot improves its performance, the value of trust will increase.

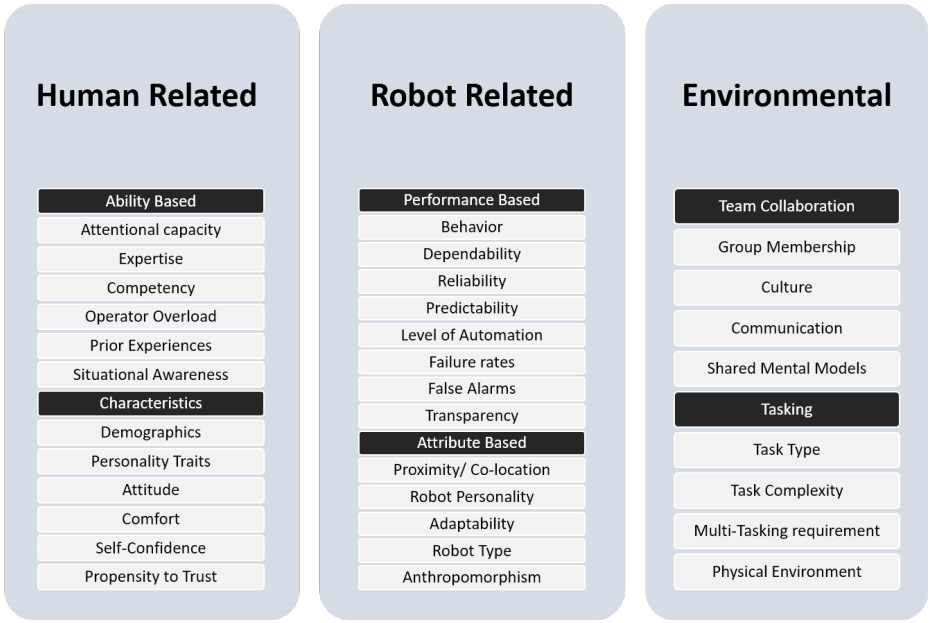

Fig. 3. Trust factors identified by Hancock [18]

Jacques Penders and his co-workers investigated HRI in 'no-visibility' conditions [34]. No visibility condition in this context means that the human user might be visually impaired, or completely blind, therefore, they would have to completely trust the robot. This study also takes a look at the attributes in the design of robots that affect the ethical behavior of them. Additionally, this study analyzes the interaction of visually impaired person and their guide dog, and examines the variables that could be implemented in the design and behavior of robots to improve the human confidence in them. These variables include human dominance, cooperation overtime, and accountability.

In [31], Stephanie Merritt at the University of Missouri examined the importance of taking differences in human behavior into consideration in the context of HRI. Ms. Merritt completed this empirical study by providing an experiment related to X-ray screening. Subjects were asked to use a simulation software to detect dangerous items such as weapons, in different luggage. They given the options of scanning the x-ray image manually, and flagging it if they spot a suspicious item, or have a fictional autonomous system called Automatic Weapons Detector(AWD) scan the image, and report any issues. This study concluded that the individual differences in subjects affects the value of trust in autonomous systems, even if the characteristics of the autonomous system is 'constant'. Thus, this study suggests that future researchers in the field of HRI and trust should take human characteristics into consideration.

Raja Parasuraman and Christopher Miller investigated the concept of trust and etiquette in the domain of HRI [33]. Given the fact that in many human-to-human social interaction scenarios, respect and etiquette highly influence the level of trust. Mr. Parasuraman and Mr. Miller argued that they also influence the perception of autonomous robots in humans. In this paper, etiquette is described as: "the set of prescribed and proscribed behaviors that permits meaning and intent to be ascribed to actions.". This study also provides an experiment related to the role of etiquette in HRI. In order to assess the influence of etiquette, test subjects used a flight simulator software called Multi-Attribute Task (MAT), and communicated with the autonomous system using different communication styles such as interrupting the user, being impatient, etc. The empirical evidence obtained by this experiment indicated that etiquette definitely influences the human trust, and reliability of autonomous robots.

A research conducted by Ms. Rosemarie Yagoda and her research partner, Mr. Douglas Gillan, provided a mechanism for measuring the value of trust in the context of HRI [51]. This measurement is based on multiple factors, including team configuration, team processes, context, task, and system. This trust scale measuring mechanism is based on two studies in this research. In this first study, subject matter experts, and previous studies were used to construct a 'content validity assessment'. The second study examines the trust scales obtained from the first study. The results of these two studies were combined to create the 'HRI Trust Measuring Tool'.

In a 2014 paper, Yue Wang and partners investigated the human-robot trust in the context of underwater semi-autonomous robots [47]. In order for the underwater robot to have a good performance, the person who is operating the robot needs to trust its autonomous capabilities. This study proposed a trust model that mainly deals with recording the robot's past performance, the human performance, and the fault rates of humans and robots. The semi-autonomous robot 'YSI EcoMapper AUV' was investigated in this study specifically. Furthermore, MATLAB simulations indicate the effectiveness of this trust model.

3.1.2 Human-Self Driving Cars (SDCs) Autonomous driving has been advancing exponentially in the recent years due to technological advancements in AI, mechanics and advanced sensor systems. Car manufacturers such as Google, Tesla, Mercedes-Benz, Ford, and many others have already created commercially available semi-autonomous cars, and fully autonomous prototypes, and they expect mass production of self driving cars (SDCs) in the early 2020s [14]. One major challenge in popularizing SDS in the US and the world would be the high level of distrust of average consumers in fully automated vehicles.

Daniel Howard at UC Berkeley (2013) [22] explores the factors affecting the trust between humans and SDCs, and attempts to examine the attitude of average consumers towards SDCs in Berkeley, California. Mr. Howard's research indicates that most consumers have positive feelings toward the ease of use that comes with SDCs. In a fully autonomous vehicle, they wouldn't have to feel frustrated when driving in heavy traffic, or finding parking in a busy area due

to the benefit of multitasking. One can imagine that at some time in the near future, commuters will be able to take naps, or watch movies while the SDC drives them to wherever they desire. Mr. Howard's also discovered that most individuals have concerns regarding the cost, liability, and the potential loss of control of SDCs. This study also indicated that factors such as level of income and gender affect the consumer's concerns. For example, subjects with higher levels of income were more concerned about liability as opposed to subjects with lower levels of income that were more concerned about the control of SDCs.

Michelle Carlson and her team, in a 2014 paper, provided a statistical analysis in the domain of autonomous vehicles and autonomous diagnostic systems [13]. Their goal was to identify the major factors that impact level of trust in self driving cars and autonomous medical systems. In this study an online survey was performed in which subjects were asked about various scenarios related to self driving cars. They were also asked about the use of IBM Watson in critical medical situations (e.g determining types of cancer). It was discovered that most test subjects were having concerns regarding the past performance of the car, reliability, errors, software/hardware failure, and the liability manufacturer of the car (e.g Google, or a lesser-known company). Similarly, it was discovered that the top factors that affect the trust in the use of IBM Watson in critical medical situations are the accuracy and past performance among many other factors. This indicates that regardless of the domain, most people tend to prioritize safety, accuracy, and failure rate when trusting an autonomous system.

In a similar study [29], Miltos Kyriakidis and his team created an international questionnaire related to the public opinion of automated driving. Questions included concerns, acceptance, and willingness to buy the car. Among the 5000 responders from 109 countries, most subjects agreed that fully automated self driving cars have the potential to be very popular among consumers by 2050. It was discovered that most subjects were concerned about safety, malicious activities/hacking, and legal issues related to autonomous driving. Mr. Kyriakidis also discovered that most of the subjects that were more educated, had more income, and were located in developed countries, were mostly uncomfortable about the self driving car transmitting data to external sources, and were concerned about the misuse of the transmitted data.

A 2015 study by Michael Wagner and Philip Koopman explored the possibilities of developing trust in Self Driving Cars using tools and techniques that are currently available [45]. This study argues that fully autonomous Driving Cars aim to make our lives easier, and reduce the number of accidents. However, in many situations such as unpredictable hazards, and intense weather situations, the human driver's reactions are superior. This is one of the major factors that affects the human trust in Self Driving Cars. It is stated that testing is a critical aspect of determining if the car is trustworthy enough to be on the road, or has the potential to develop its trust over time. This paper also claims that these cars use Machine Learning and Image Processing to provide some functions, like detecting pedestrians. It was argued, however, that Self Driving Cars need a very high accuracy in their algorithms (close to 100%), but Machine Learning algo-

rithms are not capable of producing such accurate results. Mr. Wagner and Mr. Koopman bring to light a new software testing philosophy based on philosopher, Karl Popper's concept of 'Falsification'. Falsification means: "For any hypothesis to have credence, it must be inherently disprovable before it can become accepted as a scientific hypothesis or theory."[16]. Falsifiability helps testing because it causes testers to discover more flaws and vulnerabilities in Self Driving Cars. Thus, by improving, and reliability, the trust in them will increase over time.

Tove Helldin and colleagues at the University of Skvde in Sweden conducted a study related to the ability of SDCs in snow conditions [21]. In this study, 59 drivers were chosen to sit in an autonomous simulator cockpit. One group of drivers were given information about the risks and uncertainties of the SDCs when driving in heavy snow conditions, and the other group of drivers, were not given any information regarding the ability of the SDC. This experiment indicated that the group of drivers that were provided with the information about the risks and uncertainties did not trust the SDC, and preferred to override the system to manually drive the car. The other group, however, were not as prepared to override the autonomous system to drive the car manually, and had more trust in the system.

3.1.3 Human-Autopilot Modes Scott Winter and colleagues [49] explored this low level of trust in the context of humans trusting autonomous aircraft. Subjects were asked if they preferred to be on a commercial plane with two pilots (a pilot and a copilot), a plane with a pilot in the cockpit and the copilot remotely working, or a plane with both pilots remotely controlling the aircraft. Mr. Winter and colleagues discovered that the subjects express a high degree of discomfort if they were on a fully autonomous commercial plane with both pilots just overseeing the movements and controlling the airplane remotely. They also discovered that subjects also expressed a low level of trust when only one pilot is in the cockpit and another one controlling the aircraft remotely. In this study, it was also concluded that the level of trust between humans and autonomous aircrafts is potentially related to the culture of humans. For example, it was discovered that test subjects from India feel more comfortable if they were on a fully autonomous aircraft, as opposed to subjects from the United States. Mr. Winter and colleagues found out that this difference could be due to the collectivist Indian culture, as apposed to the Individualist American culture.

Vadim Butakov and his colleague, Petros Ioannou, in [11] suggest that the level of comfort and trust of users will increase if the design and dynamics of autopilot systems in cars are closer to what they are used to in regular cars. In this study, Mr. Butakov and Mr. Ioannou analyze and present a methodology that allows custom modification of autopilot modes such as Adaptive Cruise Control (ACC) and automatic lane change systems based on individual preferences. They also support their proposed methodology by collecting data from an experimental vehicle.

At the The European Organization for the Safety of Air Navigation ([28]), a study was conducted on the human trust in air trafic management systems (ATMs), and provides guidelines and strategies to improve this trust over time. This paper argues that currently, air traffic control operators use many automated and semi-automated computer tools, and it is expected that the use of automation in this context will rise in the near future. Thus, operators will have to trust highly autonomous (or fully autonomous) ATMs such as radar systems and communication tools. The procedure to improve the mentioned trust is broken down in multiple "development phases":

- Developing ATMS using experienced air traffic controllers
- Providing high quality simulations
- Providing training for the controllers
- Transitioning period for the controllers
- Keeping the old technology in case failure

3.1.4 Human-Agents(software agents) With the fast-paced advancements in the field of IT, software agents have become very important assets that help humans with the variety of tasks such as automating repetitive computing tasks, or assisting users with simple things, such as managing their emails. There are also agents capable of doing complex tasks by themselves, without the aid of users. Examples of such agents include automatic shopping assistants, customer help desks, and web robots [43]. As these software agents grow over time, and their level of complexity increases, the issue of trust in software agents becomes one of many challenges of the use of these agents in sensitive areas. In the movie Resident Evil (2002), there is a powerful software agent called 'The Red Queen', who is in charge of a huge underground facility. Due to a bio-hazard outbreak, the Red Queen murders all the employees in the facility in order to contain the virus. Although Resident Evil is a fictional movie, it conveys an interesting concept regarding human-agent trust. In real life, Agents may become so complex that they will be able to take control of facilities or weapon systems. Thus, the level of trust in them becomes a primary concern.

In order to proceed, we need to define what an 'agent' exactly is in this context, because it is a broad term. In one major study on software agents, Dr. Hyacinth Nwana defines Software agents as computer programs that have the capability to act on someone's behalf, and the ability to make various choices by themselves [32]. In their paper 'Software Agents: An Overview', Dr. Nwana provides a detailed overview, and classification of software agents. He and his colleagues have identified three primary attributes that software agents should have: Autonomy, Learning, and cooperation. Autonomy in this case is described as "The principle that agents can operate on their own without the need for human guidance". 'cooperation' refers to the concept of multiple agents communicating, and cooperating (I will explain this in detail in the later section dedicated to multi-agent systems). The last primary attribute, 'learning" is necessary because agents need to learn, and absorb external information in order to communicate, and respond to different situations.

Dr. Nwana demonstrates that Agents also have attributes that are considered to be secondary. Attributes such as versatility (Engaging in a variety of tasks), helpful behavior, truthfulness, and the level of human-trust in them (which is what we are interested in).

Dr. Nwana identifies seven types of agents in their research:

1. Collaborative agents
2. Interface agents
3. Mobile agents
4. Information/Internet agents
5. Reactive agents
6. Hybrid agents
7. Smart Agents

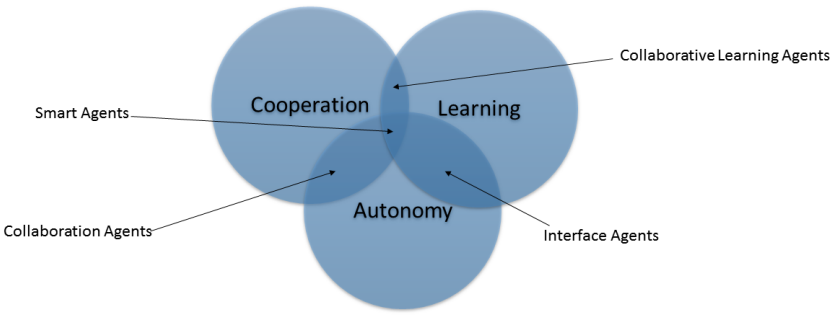

Fig. 4. Dr. Nwana's three primary attributes of agents (Dept. of IECS, Feng Chua University, R.O.C., 2003.)

In a study at MIT, Mr. Timothy Bickmore and his partner, Ms. Justine Cassell proposed an interesting model to build trust between users and computer agents [10]. Provided that in humans, trust is established and maintained by various social interactions such as small talk, Mr. Bickmore and Ms. Cassell argued that the same mechanism can also be used for computer agents to establish trust between users, and them. As an experiment, a computer agent called 'REA' was used to communicate with users by performing small talk. REA was able to track the user's movements, and detect their emotional response. The experiment in this study yielded interesting results. It was discovered that users communicating with the REA system using small talk, who also happened to

have extrovert personalities, had developed more trust than other users who had more introvert personality traits.

Mr. Bickmore's study indicates that the personality of users can potentially impact the level of trust between humans and agents. An older study conducted on the business school students in California State University yielded similar results. It was discovered that the level of self-esteem in human subjects plays a significant role in their performance and their interaction with a 'human-like' computer: "persons high in self-esteem generated more negative cognitive responses and made fewer errors when faced with human-like rather than machine-like feedback from a computer." [38]

A study at University College London titled 'Supporting Trust in Virtual Communities' proposes a reputation based trust model designed for improving trust in virtual communities. This trust model is based on 'sociological characteristics of trust', such as past experiences and reputation, or 'word of mouth'. This proposed model enables agents to take another agent's opinions and recommendations into consideration: "Our proposed model allows agents to decide which other agents opinions they trust more and allows agents to progressively tune their understanding of another agents subjective recommendations." Mr. Abdul-Rahman and his colleague, Mr. Hailes argue that although this trust model is designed for increasing trust between humans and virtual societies, this trust model can also be implemented for artificial autonomous agents, and 'artificial societies' [1].

3.1.5 Information Security At the The Norwegian University of Science and Technology, Mr. Audun Josang published a paper that aimed to focus on the concept of 'trust' in the context of information security, and shed some light on the complex nature of Trust. He also examined various types of trust, and trust driven relationships that are relevant in the context of information security, and distributed systems. Mr. Josang provides an interesting definition of trust by defining it from a malicious point of view. He explains that essentially, trust occurs due to the tendency of humans to avoid malicious behaviors of agents (Other people or software). This concept of trust and detecting malicious behavior is critical in computer security due to potential hackers or agents attempting to penetrate a system. He also explores trust and malicious behavior from a philosophical point of view. [27]

A 2006 study by Florina Almenarez et al. proposed a model for managing trust called 'Pervasive Trust Management (PTM)' [3]. 'Pervasive' devices are devices which are cumputing at all times. Apple Watches, smart-phones, PDAs etc. are examples of pervasive devices, which have microprocessors that compute things at all times in any network. The increased popularity of such devices produces the issue of trust between these devices communicating with each other, or providing a trusted, secure connection with any network. In this study, Ms. Almenarez, and her colleagues provide a statistical and mathematical trust model to tackle this problem.

3.2 Trust in Networks of Autonomous Systems

3.2.1 Multi-Agent Systems Chung-Wei Hang and colleagues provide an 'adaptive' probabilistic trust model that aims to improve the trust between autonomous agents when they interact with each other in their 2008 paper [19]. This probabilistic model extends a pre-existing model called W&S, which is primarily based on a function called PCDF (probability-certainty density function). In his proposed model, Mr. Hang provides an architecture in which autonomous agents will determine other agents' value of trust, without having any prior knowledge of them. We believe this is significant, because many of the trusted proposed trust models for autonomous systems are knowledge based. Mr. Wang, and his colleagues also developed a simulation to test their proposed trust model. Their results show that their model is highly capable of estimating the trust of autonomous agents.

'Open Distributed Systems' are described as systems that are composed of multiple autonomous agents that interact with each other based on various rules and protocols. [37]. At University of Southampton, Sarvapali Ramchurn, Dong Huynh, and Nicholas Jennings have provided a comprehensive study on the currently available trust models in the context of Multi-Agent systems [37]. They argue that most of these trust models have one fundamental goal in common: They all aim to minimize the 'uncertainty' in the interactions between agents. This study claims that currently, agents have some limitations in their capabilities due to limits in their computations and their storage devices. Thus, when interacting with other agents, or when deciding to do something, they will have a high degree of uncertainty. In 'Open distributed systems', agent would have to trust each other in order to reduce this uncertainty in their interactions.

Mr. Ramchurn and his colleagues have analyzed the trust between agents at two levels: the individual level, and system level. At the individual level, various trust models, and evolutionary models are described that help agents in Multi-Agent systems to choose their partners, and interact with them. They can also analyze their interaction with other agents to determine if they are being honest or not. At the system level, this study describes how a Multi-Agent system could be designed to enforce trustworthiness to all agents involved. An example related to agents being involved in an auction was discussed, where the increased trust in agents led to more secure bids and transactions. It was demonstrated that fear of punishment could also be an effective method: "the threat of future punishment (through avoidance of or constraining interaction(s)) could be used by reputation mechanisms to prevent agents from lying about their preferences or forcing them to behave well in an open environment"

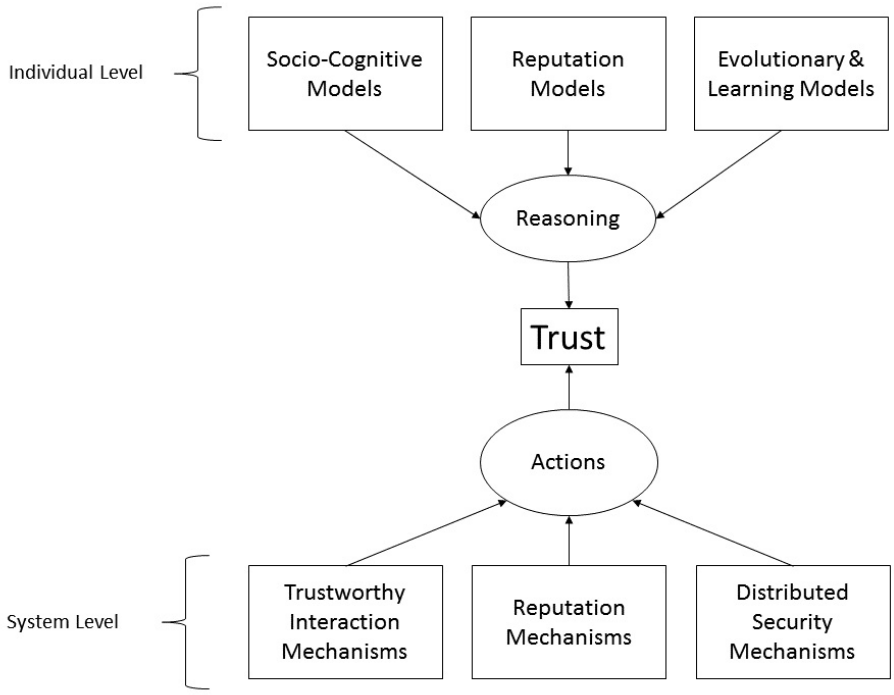

Fig. 5. Trust in Multi-Agent systems [37]

In 2009, Chung-Wei Hang et al. provided an 'Algebraic' method called 'CertProp' that aims to propagate trust between agents in Multi-Agent systems [20]. CertProp uses mathematical functions and operators such as concatenation, aggregation, and selection, in order to propagate trust between the agents. The approach that is used in this study is evidence-based, therefore, the efficiency of this approach has been proven by the evaluation of two data sets.

3.2.2 Security in Multi-Agent Networks In an article, H. Chi Wong and Katia Sycara discuss trust and security in multi-agent systems [50]. Given the fact that when in large, open systems, agents interact with other agents that they don't know, the issue of trustworthiness and security arises in these systems. This study proposes a security mechanism that could be used in such systems to enhance security and trust, and address the issues that are currently present in multi-agent systems. The paper provides methods for securing communication between agents, naming methods, and other agent services. Regarding trust between agents, it has been proposed that if entities that deploy agents become responsible for the actions of their agents, the trustworthiness of the agents will increase. Additionally, when agents are communicating, they will have to provide evidence that "They know secrets that only their delegators know."

In a paper, Yosi Mass and Onn Shehory investigated distributed trust in open Multi-Agent systems such as online trading systems [30]. They provide a new method in their study that enables agents in a Multi-Agent open system to establish trust between each other, and automatically update this trust overtime. This trust mechanism mainly utilizes 'certificates' (Asymmetric Key Exchange) in order to establish trust between agents. This paper claims that this method is efficient, and it can be easily implemented in multi-agent systems.

One paper proposed an algorithm that could be used for agents in multi-agent systems [6]. This algorithm enables the agents to analyze the incoming information from various data sources, and distinguish the reliable and trustworthy information from false, deceitful information. This algorithm uses belief systems (Bayesian Networks), which allows the agents to form beliefs based on the trust and reputation of retrieved information, and revise them overtime. This study conducted two experiments, one in static multi-agent environments, and the other in dynamic environments. The results of these experiments indicate that in static environments, this algorithm performs efficiently. In dynamic systems however, there were some delays for the agents to process and revise their Bayesian beliefs.

Another paper by Trung Dong Huynh and colleagues introduced a trust and reputation model called 'FIRE' [23]. This model could be used in open multi-agent systems to measure the performance of agents in the systems. FIRE takes into account several factors: "FIRE incorporates interaction trust, role-based trust, witness reputation, and certified reputation to provide trust metrics in most circumstances". Additionally, this study introduced a framework that aims to collect meta-data from various sources such as past interactions, witnesses, and rules about the environment for measuring the agent trust. This model allows the agents to choose their partners carefully by evaluating the trust measures they have access to. This improves the overall performance of agents in multi-agent systems.

Stefan Poslad and Monique Calisti investigated trust and security in FIPA agents, and discuss its strengths and weaknesses [35]. FIPA, which stands for 'Foundation for Intelligent Physical Agents' is a non-profit organization that was created in 1996 to produce a series of standard specifications for the development of intelligent agents in multi-agent systems. In FIPA, there are two main concepts that are discussed in this paper: agents, and agencies. FIPA agencies are described as the environment in which agents are present. They provide the agents a series with support services called 'Middle-Ware Services', which grant them the capability of interacting with other agents within the same agency, or agents from other agencies. There are two types of Middle-Ware Services: AMS (Agent management system), and DF (directory facilitator). This study examines the trust between various entities in a FIPA system. The trust model in FIPA consists of trust between agents and AMS, trust between service providers and DF, trust between service user agents and DF, and trust between agents and the 'Agent Communication Chanel'.

This study concludes that the current trust models used in FIPA systems are not optimal. As a result, it is easy for malicious agents or hackers to take advantage of service providers in order to manipulate the agents, or attack the system. Additionally, it was discovered that the current security systems related to FIPA are not effective at all, and are obsolete by today's standards.

Piotr Gmytrasiewicz and Edmunt Durfee investigated trust and honesty between fully automated autonomous agents [17]. In their paper, 'Toward a Theory of Honesty and Trust Among Communicating Autonomous Agents', they provided a method that could be used to improve communication between agents. This recursive method allows autonomous agents to calculate the outcomes of their messages before sending them to other agents. Analysis in this study indicated that although in some cases, agents might conclude the messages they are about to send might not be believable on the receiving side, communication occurs between the agents regardless. In some cases however, it was discovered that it is possible that some agents stop believing what they are told by other agents, which reduces the effectiveness of communication between them.

Y.C. Jiang et al. present a model for constructing trust between agents in multi-agent systems [26]. This 'Autonomous Trust Construction' model utilizes graph searches that allows the agents to build trust effectively. This study introduces the concepts of 'Path Searching' and 'Trust Negotiation'. Path Searching uses a graph search to find a 'Trust Path'. It is proposed that if an agent cannot find the Trust Path by performing a graph search, it can proceed to automatically 'negotiate trust' with the other agent, which mainly deals with exchanging sensitive data. The experimental results of the simulations in this study indicates that this trust construction model is highly effective.

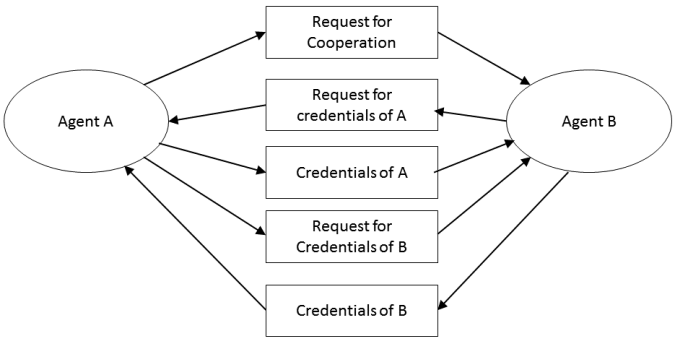

Fig. 6. Trust Negotiation example in [26]

3.2.3 Wireless Networks Wireless communications is one of the fields that has been growing very fast in the recent years. Due to increased complexity and use of wireless communication methods such as 'pervasive devices', mobile networks, and P2P networks, the need for ensuring trust and reliability in these systems arises. Mr. Tao Jiang and his research partner, John Baras provided a case study on Distributed Trust Management [25]. They introduced a statistical trust evaluation to evaluate and measure trust in agents involved in a network. Then, they provided a mathematical proof that proved that their trust evaluation method works, and affects the level of performance.

3.2.4 Social Networks Frank Walter and his colleagues argue that due to the exponential boom of the internet in the recent years, the amount of information available to uses has been increased significantly, causing 'information overload' for users [46]. Thus, there might be a need for a system that can filter the available information based on the needs of the user. As a solution, they introduced a trust-based recommendation system to be used in social networks. In this model, agents can filter the amount of information given to them based on the level of trustworthiness of other agents. They also use their social network to retrieve information that is unknown to them. Based on a series of experiments and simulations, this study indicates that the performance of this recommendation system is satisfactory.

In another similar study on trust between agents in social networks, Stefano Battiston and coworkers [7] provided an 'Automated Distributed Recommendation System' that could be implemented on agents in social networks. In this proposed model, agents use their social network in order to construct a query to conduct a search. Furthermore, it has been stated that based on the preferences of agents, the trust between the other 'neighbor' agents and the main agent could be used to filter the search result. The computer simulations in this study indicate that the 'heterogeneity' of preferences in agents in a social network plays a major role on the impact of trust on the performance of the recommendation system. It was discovered that if the agents have mild heterogeneous preferences, trust between agents positively impacts the performance of the system, as opposed to agents with higher levels of heterogeneity in their preferences.

In a study, Josep Pujol and colleagues proposed a method called NodeRanking, which could be used to extract reputation in multi-agent systems by using 'social network topology' [36]. This method is based on the idea that the position of each agent within the social network could be used in order to calculate the amount of reputation. This study also conducts an experiment on a live social network community to test its method. The paper claims that its proposed method is superior compared to other reputation extracting methods because it doesn't rely on users providing ratings (feedback) as an input.

3.2.5 P2P Networks A study by Yao Wang and Julita Vassileva proposed a 'Bayesian' networked-based trust model that could be used in P2P (peer to peer)

networks [48]. This study argues that in an open P2P network, there might be malicious, or defective agents (nodes) present in the system. This is due to the fact that in an open P2P system, there is no authority and as a result, methods of acquiring trust and reputation could be used in this systems to distinguish good peers from bad and malicious ones. This study considers a P2P file sharing system in order to explain the Bayesian method mentioned earlier. In file sharing P2P systems, there are 'seeders' (or File Provides) and 'peers' (File receivers, or agents). In this provided example, agents usually care about download speeds, file quality, and file type. Thus, agents are capable of developing a 'Bayesian' network for seeders that they have interacted with. The values of 1 and 0 are used for displaying the level of satisfaction of the agents.

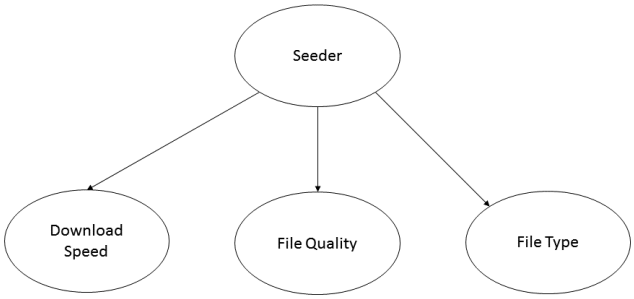

Fig. 7. Qualities that increase satisfaction in a P2P Bayesian Network

The Yang-Vassileva study tested this Bayesian trust model by simulating a file sharing P2P network, and the results indicate that agents that communicate their recommendations to each other have high performance, as opposed to agents that don't communicate as much. Ms. Wang and Ms. Vassileva claim that this approach can also be used in other areas.

4 Summary

List of Studies				
Category	Study	Summary	Technology	Focus
Human - Robots	[42]	Confidence plays an important role as well as trust, and The unpredictability of the robot affects trust	Computer simulation based on a firefighting scenario	Military, Hazardous environments
	[15]	Mathematical definition for trust, where trust is a variable T. Trust acquisition is the process or mechanism that allows the calculation and update of 'T'	trust propagation model	Math-based trust propagation. Trust acquisition
	[18]	The performance of the robot has the biggest impact on trust in the context of HRI	N/A (meta-analysis)	HRI
	[34]	In order to enhance trust in HRI, a number of design choices need to be made	visually impaired person and a guide dog	Improving trust in HRI
	[31]	Different people have different levels of trust toward robot despite its constancy	X-ray screening simulation	User perceptions of trust
	[33]	Etiquette affects human trust and the reliability of autonomous robots	Flight simulator	Impact of etiquette on trust
	[51]	Created the HRI 'Trust Measuring Tool'	Subject Matter Experts (SMEs)	Trust measurement
	[47]	If the human trusts the robot, its performance increases	YSI EcoMapper autonomous underwater robot	Semi-autonomous mutual trust
Human SDC	[22]	Most consumers like SDC's. Different people have different trust issues	Survey based on a 10 minute video	Perception of user's trust in SDCs

	[13]	Most people tend to prioritize safety, accuracy, and failure rates when trusting an autonomous system	Surveys using Survey Monkey and Amazons Mechanical Turk	Trust in Automated Cars and Medical Diagnosis Systems
	[29]	The geographic positioning and education of consumers affects the premises of their trust	Internet-based survey	User willingness to buy SDCs
	[45]	Users testing; relying and improving will increase their trust in the robot	N/A (survey of existing methods)	Safety for SDC software based on 'falsificationism'
	[21]	If people know about the problems of a robot, they will probably override it to avoid them	Driving Simulator by Volvo	The uncertainty of SDCs in various scenarios
Human-Autopilot Modes	[49]	Culture affects trust	Internet based survey	Trust in autonomous and semi-autonomous auto pilot systems
	[11]	People have an easier time trusting familiar things.	Data collection from an experimental SDC	Autopilot personalization
Human-Agents (Software)	[32]	Agents have secondary attributes that affect trust	N/A (Meta-Analysis)	Overview of software agents and comparison
	[10]	Small talk builds trust, especially with extroverted personalities. The personality of users can impact the level of trust between humans and robots.	Conversation agent REA	Building trust using social dialogue
	[1]	Model designed to increase trust between agents and agents, and humans and agents	Reputation-based trust model (Word of mouth)	Increasing trust in agents in virtual communities

Information Security	[27]	Trust occurs due to the tendency of humans to avoid malicious behaviors of agents	N/A (Analysis)	Trust in IT and malicious agents
	[3]	Mathematical trust model designed to tackle the issue of trust between agents, creating secure connections etc.	Mathematical Trust Evolution Model	Trust between agents in Pervasive devices (such as smartphones and PDAs)
Multi-Agent Systems	[19]	Model that improves trust between autonomous agents when they interact with one another	Agent simulation that uses the proposed model	interaction between agents in Multi-Agent Systems
	[37]	Agents have limitations in their capabilities, so they have a high degree of uncertainty when dealing with other agents	N/A (meta-analysis)	Factors affecting trust in open Multi-Agent Systems
	[20]	Introduces CertProp; propogates trust between agents	CertProp: a trust model	algebraic based approach in propegating trust between agents
Security in Multi-Agent System	[50]	Proposes methods for securing communication between agents	Security Architecture	Security and Trust in Multi-Agent Systems
	[6]	Propose algorithm that enables agents to tell if incoming data is trustworthy or false	Autonomous agents tracking the location of a target airplanes	Agent Belief Revision
	[30]	Provides a new method that enables agents in a multi-agent open system to establish trust between each other and automatically update the trust over time	Trust certificates	Agent trust establishment

	[35]	FIPA trust methods are not optimal and are obsolete by today's standards	N/A (Analysis)	agent standards
	[17]	Provides method that improves communication between agents	Mathematical analysis	Agent decision making
	[26]	Presents a model for constructing trust in multi-agent systems	Graph theory	Trust construction in agents
Wireless Networks	[25]	Introduces a statistical trust evaluation to evaluate and measure trust in agents involved in a network	Statistical and mathematical analysis	Trust between Agents in self-organized, autonomous networks
Social Networks	[7]	Introduces a trust-based system in which agents can filter the amount of information given to them	Computer simulations & mathematical analysis	Recommendation Systems
	[46]	Proposes a model where agents use their social network in order to construct a query to conduct a search	Multi-Agent simulations	Information filtering in agents
	[36]	proposes a method which can be used to extract reputation in multi-agent systems	Experimental ratings by NodeRaking	Trust & reputation extraction
P2P Networks	[48]	Proposes a Bayesian trust model that could be used in P2P networks	Simulation of a file sharing system	Agents in P2P networks

5 Open Problems, Future Directions and Concluding remarks

References

1. Abdul-Rahman, A., and Hailes, S. Supporting trust in virtual communities. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on* (2000), IEEE, pp. 9–pp.

2. Adams, B. D., Bruyn, L. E., and Houde, S. *Trust in Automated Systems, Literature Review.* Humansystems Incorporated, 2003.
3. Almenarez, F., Marin, A., Díaz, D., and Sanchez, J. Developing a model for trust management in pervasive devices. In *Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW'06)* (2006), IEEE, pp. 5–pp.
4. article36. ban-autonomous-armed-robots/. *article36* (2012).
5. autonomy. *Full Definition of autonomy.* merriam-webster, 2016.
6. Barber, K. S., and Kim, J. Belief revision process based on trust: Simulation experiments. In *In Proceedings of Autonomous Agents 01 Workshop on Deception, Fraud, and Trust in Agent Societies* (2001), Citeseer.
7. Battiston, S., Walter, F. E., and Schweitzer, F. Impact of trust on the performance of a recommendation system in a social network. In *Proceedings of the Workshop on Trust in Agent Societies at the Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS06)* (2006), pp. 1–15.
8. Beer, J., Fisk, A. D., and Rogers, W. A. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction 3*, 2 (2014), 74.
9. Bekey, G. A. *Autonomous Robots: From Biological Inspiration to Implementation and Control (Intelligent Robotics and Autonomous Agents).* The MIT Press, 2005.
10. Bickmore, T., and Cassell, J. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), ACM, pp. 396–403.
11. Butakov, V., and Ioannou, P. Driving autopilot with personalization feature for improved safety and comfort. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (2015), IEEE, pp. 387–393.
12. Carlson, J., and Murphy, R. R. An investigation of mml methods for fault diagnosis in mobile robots. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on* (2004), vol. 1, IEEE, pp. 180–186.
13. Carlson, M. S., Desai, M., Drury, J. L., Kwak, H., and Yanco, H. A. Identifying factors that influence trust in automated cars and medical diagnosis systems. In *AAAI Symposium on The Intersection of Robust Intelligence and Trust in Autonomous Systems* (2014), pp. 20–27.
14. driverless future. Forecasts. *http://www.driverless-future.com/6* (2016).
15. Esfandiari, B., and Chandrasekharan, S. On how agents make friends: Mechanisms for trust acquisition. In *Proceedings of the fourth workshop on deception, fraud and trust in agent societies* (2001), vol. 222.
16. explorable.com. *What is Falsifiability?* explorable.com, 2016.
17. Gmytrasiewicz, P. J., and Durfee, E. H. Toward a theory of honesty and trust among communicating autonomous agents. *Group Decision and Negotiation 2*, 3 (1993), 237–258.
18. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society 53*, 5 (2011), 517–527.
19. Hang, C.-W., , Y., and Singh, M. P. An adaptive probabilistic trust model and its evaluation. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3* (2008), International Foundation for Autonomous Agents and Multiagent Systems, pp. 1485–1488.

20. Hang, C.-W., Wang, Y., and Singh, M. P. Operators for propagating trust and their evaluation in social networks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2* (2009), International Foundation for Autonomous Agents and Multiagent Systems, pp. 1025–1032.
21. Helldin, T., Falkman, G., Riveiro, M., and Davidsson, S. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2013), ACM, pp. 210–217.
22. Howard, D., and Dai, D. Public perceptions of self-driving cars: The case of berkeley, california. In *Transportation Research Board 93rd Annual Meeting* (2014).
23. Huynh, T. D., Jennings, N. R., and Shadbolt, N. R. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems 13*, 2 (2006), 119–154.
24. INCIBE. Cybersecurity in industry 4.0. *Computer Emergency Response Team for Security and Industry (CERTSI)* (1969).
25. Jiang, T., and Baras, J. S. Trust evaluation in anarchy: A case study on autonomous networks. In *INFOCOM* (2006).
26. Jiang, Y., Xia, Z., Zhong, Y., and Zhang, S. Autonomous trust construction in multi-agent systemsa graph theory methodology. *Advances in Engineering software 36*, 2 (2005), 59–66.
27. Jøsang, A. Modelling trust in information society. *Unpublished doctoral thesis, Department of Telematics, Norwegian University of Science and Technology, Trondheim, Norway* (1998).
28. Kelly, C. Guidelines for trust in future atm systems-principles.
29. Kyriakidis, M., Happee, R., and De Winter, J. Public opinion on automated driving: results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour 32* (2015), 127–140.
30. Mass, Y., and Shehory, O. Distributed trust in open multi-agent systems. In *Trust in Cyber-societies*. Springer, 2001, pp. 159–174.
31. Merritt, S. M., and Ilgen, D. R. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society 50*, 2 (2008), 194–210.
32. Nwana, H. S. Software agents: An overview. *The knowledge engineering review 11*, 03 (1996), 205–244.
33. Parasuraman, R., and Miller, C. A. Trust and etiquette in high-criticality automated systems. *Communications of the ACM 47*, 4 (2004), 51–55.
34. Penders, J., Jones, P., Ranasinghe, A., and Nanayakara, T. Enhancing trust and confidence in human robot interaction.
35. Poslad, S., and Calisti, M. Towards improved trust and security in fipa agent platforms. In *Autonomous Agents 2000 Workshop on Deception, Fraud and Trust in Agent Societies, Spain* (2000).
36. Pujol, J. M., Sangüesa, R., and Delgado, J. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1* (2002), ACM, pp. 467–474.
37. Ramchurn, S. D., Huynh, D., and Jennings, N. R. Trust in multi-agent systems. *The Knowledge Engineering Review 19*, 01 (2004), 1–25.
38. Resnik, P. V., and Lammers, H. B. The influence of self-esteem on cognitive responses to machine-like versus human-like computer feedback. *The Journal of Social Psychology 125*, 6 (1985), 761–769.

39. Robinson, M. Science and technology in the industrial revolution. *University of Toronto Press* (1969).
40. Schaefer, K. E. *The perception and measurement of human-robot trust.* PhD thesis, University of Central Florida Orlando, Florida, 2013.
41. SINGHVI, A. Inside the self-driving tesla fatal accident. *The New York Times* (2016).
42. Stormont, D. P. Analyzing human trust of autonomous systems in hazardous environments. In *Proc. of the Human Implications of Human-Robot Interaction workshop at AAAI* (2008), pp. 27–32.
43. Tran, H. T. . T. Intelligent agent.
44. Trust. *Definition of trust in English:.* oxford-dictionaries., 2016.
45. Wagner, M., and Koopman, P. A philosophy for developing trust in self-driving cars. In *Road Vehicle Automation 2*. Springer, 2015, pp. 163–171.
46. Walter, F. E., Battiston, S., and Schweitzer, F. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems 16*, 1 (2008), 57–74.
47. Wang, Y., Shi, Z., Wang, C., and Zhang, F. Human-robot mutual trust in (semi) autonomous underwater robots. In *Cooperative Robots and Sensor Networks 2014*. Springer, 2014, pp. 115–137.
48. Wang, Y., and Vassileva, J. Bayesian network trust model in peer-to-peer networks. In *International Workshop on Agents and P2P Computing* (2003), Springer, pp. 23–34.
49. Winter, S. R., Rice, S., Mehta, R., Cremer, I., Reid, K. M., Rosser, T. G., and Moore, J. C. Indian and american consumer perceptions of cockpit configuration policy. *Journal of air transport management 42* (2015), 226–231.
50. Wong, H. C., and Sycara, K. Adding security and trust to multiagent systems. *Applied Artificial Intelligence 14*, 9 (2000), 927–941.
51. Yagoda, R. E., and Gillan, D. J. You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics 4*, 3 (2012), 235–248.