A Survey on Trust in Autonomous Systems

Shervin Shahrdar, Luiza Menezes, and Mehrdad Nojoumian
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University, Boca Raton, FL 33431
{sshahrda,lmenezes2014,mnojoumian}@fau.edu

Abstract—**As a result of the exponential growth in technology and computing in recent years, autonomous systems are becoming more relevant in our daily lives. As these systems evolve and become more complex, the notion of trust in human-autonomy interaction becomes a prominent issue that affects the performance of human-autonomy teaming. Prior studies indicate that humans have low levels of trust in semi and fully autonomous systems. In this survey, we review a wide range of technical papers and articles and go over the related experimental techniques in the literature of trust. We also explain limitations that are present in existing research works, and discuss open problems in this domain. It's apparent that trust management is critical for the development of future artificial intelligence technologies.**

Keywords*—*Trust management, autonomous systems; human-robot interaction; self-driving cars; autopilot systems.

I. Introduction

The rapid growth of technology has resulted in automation of many daily tasks humans had to perform themselves just decades ago. From industrial robots used in factories to autopilot systems, automation continues to aid humans with repetitive, tedious, and monotonous tasks. Every day, automation introduces newer and more sophisticated concepts and automated systems in different areas of our lives including, but not limited to, our homes, daily commute, workplace, etc.

As automated systems evolve and levels of complexity rise overtime, their presence in various daily activities of humans leads to the development of new notions in human-autonomy interaction, e.g., trust, satisfaction, frustration, to name a few. Studies have indicated that one of the primary challenges for successful integration of advanced autonomous systems and artificial intelligence technologies in human civilization will be the management and development of mutual trust [1].

The possible misuses and abuses that humans would bring into automation technologies is another prominent issue in this domain. According to [2], users can become overly dependent on an automation technology, attempt to use functions that are out of the scope of the system, or not monitor the system properly. These are due to trust or distrust in these technologies. This highlights the importance of educating those who intend to purchase an autonomous technology, mainly with respect to proper use of these technologies.

Furthermore, the usage of information that is given to users with respect to an autonomous system is also important. In [2], human subjects were given information sources regarding usage of an autonomous system with faulty behavior. This study revealed that, when more errors occur, the participants will not use the information provided to them due to lack of trust. Therefore, a proper trust management approach helps the users to utilize such information even in the presence of errors.

Finally, consideration should be given to safety of autonomous systems. Indeed, users tend to be unaware of functions that an autonomous vehicle is capable of carrying out, often due to the over complexity that appears throughout the system. The authors in [3] state that safety would increase with simplification and possible training through an interface between autonomous systems and drivers.

We are therefore interested to further discuss on trust in autonomous systems. In section II, we will discuss the definitions of trust and autonomous systems. In section III, we will provide a comprehensive analysis and explore the related literature. We will go over experimental techniques proposed by researchers in this field to understand human-autonomy trust management. Finally, in section IV, we will go through the limitations in the current literature and provide our concluding remarks.

II. Basic Definitions

A. Autonomous Systems

The definition of an autonomous system continuously changes [4]. 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, and politics. It is worth mentioning that, the second part of the autonomous term, i.e., *nomos*, means *law* in Greek. Therefore, an autonomous system is an entity that creates its own laws [5]. A more specific definition would be, intelligent machines that are capable of performing tasks by themselves and without explicit human control [6].

Although automation was introduced to human civilization many years ago and it is widely utilized after the industrial revolution [7], autonomous systems and industrial robots are relatively new technologies that were introduced merely decades ago. In this survey, our primary focus will be on autonomous and semi-autonomous robots or machines, self-driving cars, and autopilot systems. These are highly advanced forms of automated technologies. Furthermore, these autonomous systems have high levels of self-awareness and are capable of independently performing various tasks.

B. Trust and Its Measurement

Trust is a term that has many different definitions in various contexts such as psychology, sociology, economics, and computer science. Currently, there is no uniform definition of trust [8]. Prior research indicates that there are over 300 definitions in various research areas, and in the context of human-autonomy interactions, there are too many definitions and notions of trust. These notions include measurement of trust [4], computational models of trust [9], and human-inspired models of trust [10], to name a few. The formal definition of computational trust can be defined as follows [11]:

Definition 1: Let $T_i^j(t)$ be the trust value *assigned by P_j to P_i in period t. Let $T_i(t) : \mathbb{N} \mapsto \mathbb{R}$ be the trust function that illustrates how trustworthy P_i is, i.e., a mapping from natural (cooperative or non-cooperative action) to real numbers:*

$$T_i(t) = \frac{1}{n-1} \sum_{j \neq i}^{n} T_i^j(t),$$

where $-1 \leq T_i(t) \leq +1$ defines the upper and lower bounds of trust value, and $T_i(0) = 0$ determines the initial trust value, which is assigned when interaction starts.

Indeed, this formal definition illustrates that trust can be quantified and measured. Accordingly, actions or behaviors can be modified. The ultimate goal in technological systems is to have a quantifiable model of trust so that the system can be responsive to human trust.

Oxford dictionary defines the notion of trust as "A firm belief in the reliability, truth, or ability of someone or something [12]." If we consider this simple definition, in our context, the definition of trust would be: a strong human belief in the reliability, truth, or ability of an autonomous system. Some examples of trust in autonomous systems include trust in robots, machines, self-driving cars, autonomous airplanes and software agents.

III. Trust Between Humans and Autonomous Systems

In this section, we review the related studies in four categories in order of publication date. The classification includes trust in human-robot or human-machine interaction, trust in self-driving cars, and trust in autopilot systems.

A. Trust in Human-Machine Interaction

Muir [13] provided an analysis on trust in human-machine interaction from a psychological perspective. The author illustrated how trust is evolved when *Decision Support Systems* (DSS) are used. DSS are computer programs that assist an individual or an organization when making decisions. These decisions include ranking important documents, buying and selling stocks quickly, choosing a target market, and many other important decisions [14]. Since DSS have a high impact on some critical decisions, the user's trust in such systems becomes a crucial factor when designing decision support systems. The author initially analyzed the psychological trust models among humans and then developed a human-machine trust model. In this model, the concept of *trust calibration* is introduced in which the user has the responsibility of calibrating their trust based on the reliability of the decision support system. The author suggested that there are certain factors and design goals that should be considered in order to better calibrate trust when designing decision support systems. These factors are as follows:

1) *Aiming to improve the user's perception of trustworthiness of decision support systems*: This would require the users to understand how a DSS works, and also, to become familiar with the predictability of the system's decisions. The author recommended that this can be achieved by putting each user on a trial period of using a DSS to improve the user's perception. A candidate solution would be the use of a simulation environment so that the user can freely explore the DSS without any fear or concern related to wrong or dangerous decisions.
2) *Modifying the decision support systems' criterion of trustworthiness*: In order to achieve this, the DSS has to provide a history of efficiency and good performance. It was recommended that the users have access to statistical data such as the system performance.
3) *Continuous identification and fixing the causes of poor trust calibration*: In order to improve the trust between humans and machines, the system (or developers) should detect bad trust calibrations and fix them. This study indicated that some of the main causes of low trust might be due to incorrect expectations of the users. Thus, the calibration training for the users is crucial.

Another fascinating study, which examined the role of trust in decision support systems and autonomous aids, was conducted by Madhaven and Wiegmann [15]. Since the role of DSSs and autonomous aids for making critical decisions has been significantly increased overtime, this paper proposed a framework by which human trust in autonomous aids can be increased overtime. The proposed trust framework utilizes the psychological traits that affect trust among humans. It uses these traits to provide a set of instructions for the DSS so that human trust is increased. The outcome of this research contributed to the identification of several important psychological factors such as favoritism (human vs. robot partners) and subjective bias in users, which affect human-robot trust relationship. These variables are critical for the development of decision support systems as well as autonomous systems that interact with humans.

Factory automation is another line of research in this domain. Lee and Moray [16] executed an experiment to characterize variation of operator's trust during an interaction with a semi-automatic pasteurization plant. The authors investigated the relationship between changes in operators' control strategies and trust. In the same line of research and based on [17], Muir [18] conducted two experiments to test the influential variables in human-machine trust, and provided experimental analysis on the theoretical trust model proposed a few years earlier. The results of these experiments indicated that the perception of competence of an autonomous system relates to the amount of trust a user may have in the system. For example, if a user detected that the system might be incapable of doing its job, i.e., incompetence, they would manually take control of the system, as a result, their trust would drastically decrease. Another finding of this study indicated that the amount of monitoring the autonomous system will decrease if the level of trust in the system increases. The author suggested that the findings in this research could be used by industry professionals to determine which properties of autonomous systems could have vulnerabilities that might display incompetence and lower human trust. By doing that and predicting the patterns of human trust, they would be able to increase the overall effectiveness of the autonomous system.

Dassonville et. al. [19] investigated the issue of trust within the context of a teleoperation system, which is a type of system in which human operators control a machine/system from a distance. The authors first analyzed the role of trust in human relationships and then extended this study to human-machine systems. They also conducted experiments on the role of self-confidence in human-machine interactions. A teleoperation system is simply composed of three components as follows:

1) *Master universe*: The master universe is the environment in which the operator resides in. An example of this would be a military drone operator sitting in a container in the middle of a desert and controlling a strike drone somewhere far away in a combat zone.
2) *Slave universe*: Similarly, the slave universe is the environment in which the machinery or the system operates through the operator's commands. This universe is composed of hardware and a group of sensors.
3) *Space between the master universe and the slave universe*: The space between the master universe and the slave universe contains data transmission (e.g. Internet), fast computers, and decision control systems.

In this study, a simulated experiment was performed by having an operator to use a joystick (master universe) that was connected to a computer (space in between) to control a cursor on the screen (slave universe). This experiment was conducted on two student populations of literature studies and scientific studies. The study discovered that the first population appeared more self-confident in operating the machine, however, both groups had similar levels of trust in the system.

In [20], Moray and Idnagaekri analyzed how trust in autonomous systems leads to a lower level of human supervision. They discovered that as human subjects become more reliant on the system, a decrease in constant monitoring occurs. This study also scrutinized the situation in which participants overly trusted a targeted autonomous system, which could lead to malpractice in some cases.

The authors in [21] conducted a study based on effects of continuous and discrete malfunctions within autonomous systems. The study had two parts. The first section tested human participants based on continuous and discrete faults separately, whereas the second part intertwined two malfunction types. These experiments found a significant decrease in trust after five continuous failures. However, there was no significant reduction in trust after one discrete malfunction. The results of this study demonstrated how trust was dissipated and how users relied on autonomous systems based on previous faults.

Dzindolet et. al. [22] performed an experiment to improve trust in autonomous systems. This study involved participants detecting a soldier camouflaged in an area. They had the option of manually guessing whether the soldier was there, or have assistance from an automated aid. The first study measured human trust in the system before any interaction with the system. The results indicated that the operator would trust the system that had higher approval ratings and fewer errors. The second study compared the number of mistakes the user made to the number of errors produced by the system. To accomplish this, two separate groups were selected. One group had a system that made twice as many errors as the user, and the other system made half as many errors as the user. The result demonstrated that those who had more errors were more inclined to stick to their decision.

Finally, Merritt [23] examined the importance of considering differences in human behaviors in the context of human-automation interaction. The author conducted an 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 luggage. They were given the options of scanning the x-ray image manually and flagging it if they spot anything suspicious, or have a fictional autonomous system, called *Automatic Weapons Detector* (AWD), to examine the image and, consequently, report any issues. This study found that the individual differences in subjects affect the value of trust in autonomous systems, even if the characteristics of the autonomous system is constant. This study suggests that future researchers should consider human characteristics when designing experiments for trust analyses and measurements.

The summary of this section's results is shown in Table I.

B. Trust in Human-Robot Interaction

Murphy et. al. [24] investigated the use of autonomous rescue robots in combat situations as well as cases where victims were unable to be reached, for instance, victims stuck in earthquake rubble. In this study, the purpose of the rescue mission was to find the victims, check for vital signs, and help the victims until they are rescued. The study discovered that the success of these robots simply depend on victims' trust. In other words, it was crucial that the victims allow the robot to help and collect data for an optimal recovery and assistance, which could be achieved if a high level of trust was established.

Human trust also depends on the failure rate of the autonomous systems. For instance, a study of commercially available ruggedized robots operating under field conditions showed a *Mean-Time-Between-Failures* (MTBF) of 12.26 hours and an availability rate of 37% [25]. This finding indicates that if the robotic systems reduce their failure rates, their reliability will increase, and subsequently, the confidence in their performance will increase. It is apparent that, due to recent advancements in technology, the mean-time-between-failures has been decreased in autonomous systems even outside of the robotics.

Parasuraman and Miller [26] investigated the concept of trust and etiquette in the domain of *Human-Robot Interaction* (HRI). Given that respect and etiquette highly affect the level of trust in many human-to-human social interaction scenarios, the authors argued that these factors also have impacts on perception of humans with respect to autonomous robots. In this study, etiquette is described as a set of prescribed and proscribed behaviors that permits meaning and intent to be ascribed to actions. This study also conducted an experiment related to the role of etiquette in HRI. Human 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 showed that etiquette affects human trust as well as the reliability of autonomous robots.

TABLE I. TRUST IN HUMAN-MACHINE INTERACTION

Reference	Summary	Approach	Concentration
[13]	User has the responsibility of calibrating their trust based on the reliability of the decision support system	Analytical	Trust calibration in decision support system
[15]	Introduced an advanced framework to improve trust overtime	Trust framework	Decision support systems
[16]	Changes in operators' control strategies and their connections to trust	Simulated semi-automatic pasteurization plant	Control strategies for autonomous systems
[18]	If a user detects the system might be incapable of doing its job, they would manually take control of the system	Simulated semi-automatic pasteurization plant	Incompetence of autonomous systems
[19]	Investigated the role of self-confidence and trust in teleoperation systems	A joystick and a computer	Trust and teleoperation systems
[20]	Theoretical trust models and past experiments were explored and summarized	Survey-based	Human-machine interaction
[21]	Demonstrated the changes in trust based on past failures	Simulated pasteurization plant	Malfunctioning and its impact on trust
[22]	Trust is a critical factor in automation reliance decisions	Slides and graphs were shown to test human subjects	Automation reliance
[23]	Different people have various levels of trust toward an autonomous system despite of its constancy	X-ray screening simulation	User perceptions of trust

Fig. 1. Autonomous and semi-autonomous robots used in battlefields [28].

Stormont [27] showed a low level of trust in HRI. The author investigated the factors that affect trust between humans and robotic systems. He discovered that one of the reasons for such a low level of confidence in autonomous systems is due to their low level of reliability. The author also discovered that unpredictability is another factor affecting trust between humans and autonomous systems. He argued that, in various hazardous circumstances such as battlefields - as shown in Figure 1 - and rescue missions, the unpredictability of robots becomes a significant problem for human supervisors. Although the autonomous nature of robots and their quick decision making are known as positive traits, the problem arises when life and death of humans will depend on the choices of a robot. Indeed, questions such as “Should life and death decisions be made by an autonomous system?” have been in the center of attention by many researchers. The same study executed a simulation of robots assisting firefighters in a hazardous fire situation. The simulation showed that, even though firefighters did not initially trust these robots, their reliance and trust in the robots increased as the mission progressed and they became tired. As a result, they finally let the robots to extinguish the fire.

The authors in [29] investigated how culture and appearance might have an impact on trust. The team sampled participants from China, Germany, and Korea to analyze different cultural backgrounds. The participants were asked to interact with a robot that knew the culture of participants' countries. The results were scaled based on likeability, engagement, trust, and satisfaction. The outcomes demonstrated that the robot appeared differently to each participants, thus showing the need for increased concentration in different areas. Producers may use this information to create their robots more unique to specific regions and cultures to improve trust and demand within the communities.

Hancock et. al. [30] provided a comprehensive analysis of factors affecting trust in human-robot interaction. This study classified factors affecting trust in HRI into three different categories, i.e., human, robot, and environment, as shown in Figure 2. Human-related factors include training, expertise, situational awareness, and demographic information. Similarly, robot-related factors are behavior, dependability, reliability, level of automation, failure rates, false alarms, transparency, and attribute-based factors such as location, personality, adaptability, robot type, and anthropomorphism (having human traits). Finally, environmental factors include teamwork, culture, communication, shared mental models, task type, task complexity, and multi-tasking. This paper discovered that robot performance has the highest impact on human trust.

Yagoda and Gillan [31] proposed a new mechanism for measuring the value of trust in the context of HRI. This measurement was based on multiple factors such as team configuration, team processes, context, task, and system. The proposed trust measuring mechanism was developed using two experiments. The results of these two studies were combined to create a new HRI trust measuring tool.

Penders et. al. [32] investigated HRI in “no-visibility” conditions, which means the human subjects might be visually impaired or blind. Therefore, they would have to trust the robot completely. This study analyzed the interactions of visually impaired people with their guide-dogs and examined the variables that could be utilized in the design and behavior of robots for improvement of human trust. These variables include human dominance, cooperation overtime, and accountability. It is worth mentioning that Castelfranchi and Falcone [33] investigated how “control” affects trust negatively. Their research discovered that human trust will decrease if a participant is forced to take control of an autonomous system. We do believe that this is a prominent issue that should be considered in no-visibility conditions.

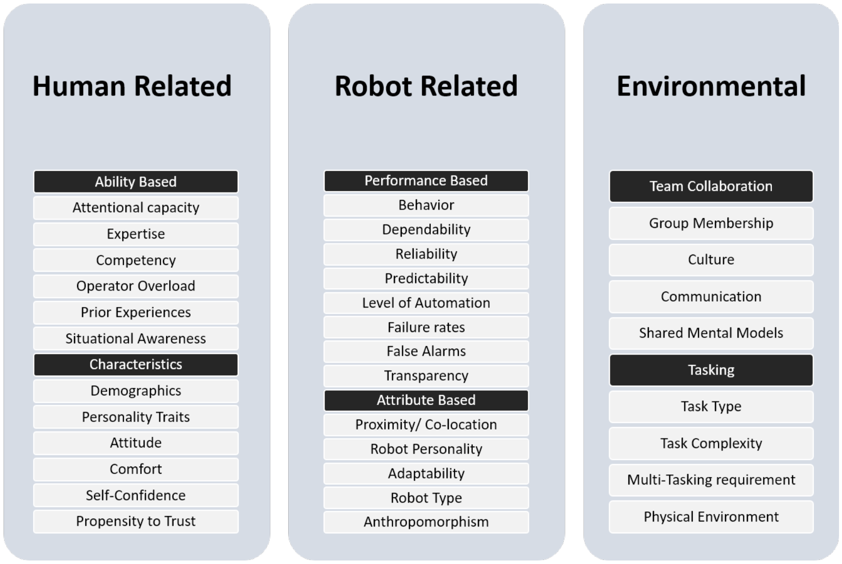

Fig. 2. Trust factors identified by Hancock [30].

Wang et. al. [34] investigated the human-robot trust in the context of underwater semi-autonomous robots. To increase the performance of a submarine robot, the operator's trust in capabilities of the submarine robot must be established and sustained. 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. A semi-autonomous robot, known as YSI EcoMapper AUV, was tested in this study. The authors show the effectiveness of this trust model through a simulation-based approach.

An essential aspect of improving human trust with respect to robotic systems is the real-time measurement of the emotional response of human subjects. If the artificial intelligence technologies can learn and interpret the emotional states of human, they will be able to adjust themselves accordingly to be responsive to those emotional states. Hu et. al. [35] provided a trust sensor model that utilizes psychophysiological measurements of human subjects. The objective of this project was to find out if psychological factors of humans captured through sensors, such as *Electroencephalography* (EEG) and *Galvanic Skin Response* (GSR), can be used to manage trust in the context of human-robot interaction. After a series of experimental studies, statistical analyses and classification, this study concluded that psychophysiological measurements could be used to measure trust in humans. However, the mean accuracy of this claim was 71.57%. Therefore, this method cannot be used for all humans. The authors believe that consideration should be give to human subjects' demographic information in any experimental study.

Finally, due to technological advancement in recent years, care-giving robots are becoming popular and common in our lives. As a result, trust in these robotic systems is a prominent issue from consumers' perspectives. In a recent study [36], [37], the authors measured the levels of trust, satisfaction, and frustration in the context of care-giving robots by designing several experiments in which human subjects interacted with a Baxter robot through a sequence of trust-building or trust-damaging incidents. In this study, various scenarios were tested, for example, delivering an object with different speeds, accidentally dropping the object, etc. The authors discovered that performance of the assistant robot affects the human trust, satisfaction and frustration when a sequenced of structured incidents are considered.

The summary of this section's results is shown in Table II.

C. Trust in Self Driving Cars

Autonomous driving has been advancing rapidly in the recent years due to technological advancements in software and hardware platforms, artificial intelligence, and sensor and radar systems. Car manufacturers along with high tech companies (e.g., Tesla, Mercedes-Benz, BMW, Porsche, Volvo, Ford, Waymo, Uber, etc) have already created commercially available semi-autonomous cars and fully autonomous prototypes. They intend to mass produce *Self-Driving Cars* (SDC) in early 2020s [38]. One major challenge in popularizing self-driving cars in the US and the world is the average consumers' high level of distrust in fully automated vehicles.

Uggirala et. al. [39] analyzed trust in a situation in which the users were given information about the capacity of the autonomous vehicle. This study aimed to decrease uncertainty to optimize system performance by having the users to be knowledgeable about the functions that the autonomous car can perform. The participants went through training to become familiar with the system. Subsequently, they had to judge whether or not the vehicle would be capable of efficiently completing certain functions given reference lines. This study concluded that, when users are knowledgeable about the system, their trust in the system increases.

The authors in [40] conducted a study related to the ability of self-driving cars in snow conditions. 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 SDC when driving in heavy snow conditions, whereas the other group of drivers did not know anything about the ability of the SDC. This experiment indicated that the group of drivers who were knowledgeable about the risks and uncertainties did not trust the SDC and preferred to override the system in order to drive the car manually. However, the other group didn't override the autonomous system to drive the car manually and had more trust in the system.

In [41], Howard focused on factors affecting trust in self-driving cars. The author examined the attitude of average consumers towards SDCs. This research discovered that most consumers have positive feelings toward the ease of use that comes with self-driving cars. In the context of fully autonomous vehicles, users wouldn't have to feel frustrated when driving in heavy traffic or finding parking in busy areas. One can imagine that, at some time in the future, commuters will be able to take naps or watch movies while the car drives them to wherever they desire. The author also discovered that most individuals have concerns regarding the cost, liability, and the potential loss of control in SDCs. Income and gender are other variables that affect the consumer attitude towards SDCs. For example, subjects with higher levels of income were more concerned about liability, but subjects with lower levels of income were more concerned about loss of control.

Carlson et. al. [42] conducted a statistical analysis in the domain of autonomous vehicles and autonomous diagnostic systems. They created an online survey and asked human subjects about various scenarios related to self-driving cars and usage of IBM Watson in critical medical situations (e.g., to determine types of cancer). It was discovered that most test subjects had concerns regarding the past performance of the car, reliability, errors, software/hardware failures, and the liability of the car manufacturer. Similarly, it was discovered that top factors that affect trust during the usage of IBM Watson in critical medical situations are accuracy and past performance. The result of this study indicated that, regardless of the domain, most people tend to prioritize safety, efficiency, and failure rates when deciding to trust an autonomous system.

Kyriakidis et. al. [43] created an international questionnaire related to the public opinion of automated driving. Questions consisted of concerns, acceptance, and willingness to purchase a self-driving car. Among 5000 participants from 109 countries, most subjects agreed that fully automated cars have the potential to be very popular among consumers by 2050. However, the majority of subjects were concerned about safety, malicious activities/hacking, and legal issues related to autonomous vehicles. The authors also found that most of the educated subjects had more income and were located in developed countries. This class of human subjects were uncomfortable with the self-driving cars transmitting data to external sources. They were also concerned about the malicious usage of the transmitted data.

The authors in [44] explored the possibilities of developing trust in self-driving cars using techniques that are currently available. This study argued that SDCs aim to make our lives easier and reduce the number of accidents. However, in many situations such as unpredictable hazards and intense weather situations, human driver's reactions are superior and needed. Testing was discovered to be a critical aspect of determining if the car is trustworthy enough to be on the road or has the potential to develop its trust overtime. Self-driving cars utilize machine learning and image processing techniques to provide functions like detection of pedestrians or stop signs. It was argued that many people require a very high accuracy in the functions of their SDCs (close to 100%), but machine learning algorithms cannot produce such accurate results.

Butakov and Ioannou [45] suggested that the levels of comfort and trust of users will increase if the design and dynamics of autopilot systems in cars are closer to what they are in regular vehicles. In this study, the authors analyzed and presented a methodology that allows custom modification of autopilot modes such as adaptive cruise control and automatic lane change systems based on individual preferences.

TABLE II. TRUST IN HUMAN-ROBOT INTERACTION

Reference	Summary	Approach	Concentration
[24]	Success of the rescue robots depend on trust of the victim whom they are assisting	Survey-based	Rescue robots
[25]	If robotic systems reduce their failure rates, their reliability will increase	Simulator based on robot behavior	Improving trust by reducing failure rates
[26]	Etiquette affects human trust as well as the reliability of autonomous robots	Flight simulator	Impact of etiquette on trust
[27]	Unpredictability of robots affects trust	Computer simulation based on a fire-fighting scenario	Military and hazardous environments
[29]	The appearance of a robot affects human perception and trust	Lego Mindstorm NXT robot programmed	Trust and cultural differences
[30]	The performance of robots has the highest impact on trust in the context of human-robot interaction	Survey-based	Human-robot interaction
[31]	Created a trust measuring tool for human-robot interaction	Subject matter experts	Trust measurement
[32]	To enhance trust in human-robot interaction, a number of design choices need to be made	Visually impaired person and a guide dog	Improving trust in human-robot interaction
[33]	Control on autonomous systems increases or decreases trust depending on the circumstances	Cognitive analysis	Trust and artificial intelligence
[34]	If humans trust a robot, its performance will increase	YSI EcoMapper autonomous underwater robot	Trust and semi-autonomous robots
[35]	Psychophysiological measurements can be used by artificial intelligence to measure human trust	Human subject study involving 31 humans using EEG and GSR	Psychophysiological measurements and trust
[36][37]	Performance of a robot directly affects trust, satisfaction and frustration	Human subject study involving 10 subjects using an assistant robot	Trust, satisfaction, and frustration

Payre et. al. [46] conducted an experiment and analyzed how alternating levels of trust would affect a driver's reaction time. The main objective of this experiment was to measure time for the *Manual Control Recovery* (MCR) when emergency situations arise in fully autonomous cars. These cars are certified and eligible to be used by drivers with a standard driver license. Nonetheless, the drivers are still accountable for their vehicles, regaining MCR in the case of emergency, and being in the driver's seat with their seat-belt buckled at all times. This study demonstrated that higher trust levels in fully autonomous cars resulted in slower reaction times, which can create a hazardous environment for drivers. This experiment can help companies become more aware of problems that over-trusting can cause.

Akash et. al. [47] presented a gray-box modeling approach that has the capability of capturing different variations of human behavior related to human-machine trust. In an experiment involving 581 human subjects, the authors utilized a computer simulation platform by which human subjects were shown an obstacle detection system (as used in self-driving cars) and were asked if they trust or distrust the image processing algorithm used in the system. The study discovered that human trust significantly decreases in faulty scenarios. An observation of this study was that the amount of trust in the system slightly increases after around 8 to 10 trials when a negative experience has already lowered their trust. Additionally, this study investigated the effects of national origins, culture, and gender on the level of trust. The results indicated that Americans usually have less trust in autonomous vehicles compared to people from other nationalities such as Mexicans and Indians. This finding matches previous discoveries in this domain. The overall conclusion of this study suggests that a perfect autonomous system should be able to collect data (e.g., psychological factors, demographic information, etc.) from its users and use that data to maintain and improve trust.

Finally, Daziano et. al. [48] investigated the willingness of consumers to purchase self-driving cars by conducting an online experiment over 1260 human subjects all around the world. After analyzing the data, it was estimated that the average household tends to pay about $\$3,500$ more for partial automation, and $\$4,900$ for full automation when purchasing a new vehicle. This research also found that the preferences of consumers with regards to different levels of automation (i.e., low, semi, full) highly varies. A significant portion of participants even preferred to pay more than $\$10,000$ extra for fully automated vehicles. Based on the information from this study, we can observe a pattern in consumers' behavior, which suggests that the public interest in self-driving cars is increasing rapidly. We believe that this public interest will spike in the near future as issues such as trust between humans and self-driving cars as well as reliability of autonomous technologies are resolved.

The summary of this section's results is shown in Table III.

D. Trust in Autopilot Systems

In the final part of our survey, we mainly focus on human trust in systems with autopilot capabilities.

de Vries et. al. [49] showed how planned routes in manual and automatic modes affected trust. This experiment had a group of participants to plan a route and then choose to complete it manually and automatically ten times each. The study proved that automatic failures had more negative impacts on trust compared to manual failures. Participants were more likely to forgive themselves for the error they had committed than the failures that happened during automatic mode. The results demonstrated a bias towards participants trust in manual mode as opposed to automatic mode.

In [50], the authors conducted research on human trust with respect to air traffic management systems. They provided guidelines and strategies for improving trust in autopilot systems overtime. They argued that air traffic control operators currently utilize many automated and semi-automated computer tools and more usage of fully automated systems is anticipated in the near future. Thus, the operators will have to trust components of autonomous (or fully autonomous) air traffic management systems, for instance, radar systems and communication tools. The procedure to improve the aforementioned issue can be formed through multiple development phases as follows:

1) Developing systems by experienced air traffic controllers.
2) Providing high-quality simulations.
3) Providing training for the controllers.
4) Transitioning period for the controllers.
5) Keeping the old technology for the case of failures.

Jiang et. al. [51] discovered that there is a direct correlation between the specific type of errors that occurs and operators' trust in the autonomous system. In this study, the participants were monitored throughout a week. The first day was solely based on training to recognize errors and functions of the system. During the rest of the trial, the team examined how participants felt about false alarms, given by high-risk systems. The results demonstrated that there was a significantly greater decrease in trust towards systems that continuously outputted false alarms.

In an experimental study [52], the authors investigated the methods in which human subjects were able to judge the performance of complex autonomous systems. To accurately investigate this, the participants were put through training that would train them on measuring the accuracy and performance of airplanes. The participants were then asked to analyze an airplane's performance and rate it as friendly or hostile based on the measured speed, altitude, range, and time in the air. The results of the study were relatively accurate post training, and they demonstrated that the judgments became accurate when participants learned what they were looking for in complex autonomous systems.

Finally, Winter et. al. [53] investigated distrust of humans with respect to autonomous airplanes. In this study, human subjects were asked if they prefer to be on a commercial airplane with two pilots (a pilot and a co-pilot), an airplane with a pilot in the cockpit and a co-pilot working remotely, or an airplane with both pilots controlling the aircraft remotely. The authors expressed that the human subjects would have a high degree of discomfort if they were on a fully autonomous commercial airplane with both pilots just overseeing the movements and controlling the airplane remotely. They also mentioned that the subjects would have a high degree of distrust when only one

TABLE III. Trust in Self-Driving Cars

Reference	Summary	Approach	Concentration
[39]	When users are knowledgeable about an autonomous system, their trust in the system increases	The participants went through training to become familiar with the system	Training and education
[40]	Drivers who are knowledgeable about risks, when driving in snow conditions, do not trust self-driving cars	An autonomous simulator cockpit	Knowledge about risks and uncertainties
[41]	Consumers have positive feelings toward the ease of use that comes with self-driving cars	Survey-based	Factors affecting trust in self-driving cars such as gender and income
[42]	Past performance, reliability, errors, software and hardware failures will affect trust	Online survey	Impacts of safety, efficiency, and failure rates on trust
[43]	Self-driving cars will be popular, however, users are concerned about safety, hacking, and legal issues	Survey-based	The future of self-driving cars and major concerns
[44]	Unpredictable hazards are still an issue that needs to be resolved	N/A	Self-driving cars and safety
[45]	Level of trust increases if the design and dynamics of SDC are closer to what they are in regular vehicles	Data collection from an experimental self-driving cars	Autopilot personalization
[46]	Over-trust is an issue and can potentially cause hazardous situations	Visual channels and 10 screens in order to analyze reaction time of users	Educating consumers on the proper time to regain manual control over vehicle
[47]	A model to capture the dynamic variations of human trust	Experiment involving 581 subjects	Dynamic trust and impacts of demographic information
[48]	Consumers are willing to pay significantly more for autonomous features	Experiment involving 1260 subjects	Consumer behavior

pilot was in the cockpit. This study also discovered that the human trust in autonomous airplanes is related to the culture of humans. For example, the test subjects from India felt more comfortable if they were on a fully autonomous aircraft as opposed to subjects from the United States. The authors found that this difference could be due to the collectivist Indian culture as opposed to the Individualist American culture.

The summary of this section's results is shown in Table IV.

IV. Conclusion and Future Direction

In this survey, we thoroughly reviewed the existing literature of trust in autonomous systems. We went over technical papers/articles that examined trust between humans and robots, machines, self-driving cars, and autopilot systems. Many of the reviewed studies provide new discoveries as well as recommendations to manage and improve trust between humans and fully or semi-autonomous systems. The literature on trust, however, is still very broad and does not address concrete trust issues that are currently present, for instance, how all existing discoveries can be translated to computational trust models that are understandable to machines. New research directions and novel methodologies can potentially provide a solid platform to develop well-performing autonomous systems.

V. Acknowledgment

We would like to thank FAU for supporting this project.

References

[1] J. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of Human-Robot Interaction*, vol. 3, no. 2, p. 74, 2014.

[2] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, 1997.

[3] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[4] K. E. Schaefer, *The perception and measurement of human-robot trust*. PhD thesis, University of Central Florida Orlando, Florida, 2013.

[5] autonomy., *Full Definition of autonomy*. merriam-webster, 2016.

[6] G. A. Bekey, *Autonomous Robots: From Biological Inspiration to Implementation and Control (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[7] M. Robinson, "Science and technology in the industrial revolution," *University of Toronto Press*, 1969.

[8] B. D. Adams, L. E. Bruyn, and S. Houde, *Trust in Automated Systems, Literature Review*. Humansystems Incorporated, 2003.

[9] M. Nojoumian and T. C. Lethbridge, "A new approach for the trust calculation in social networks," in *E-business and Telecommunication Networks: 3rd International Conference on E-Business, Best Papers*, vol. 9 of *CCIS*, pp. 64–77, Springer, 2008.

[10] M. Nojoumian, "Trust, influence and reputation management based on human reasoning," in *4th AAAI Workshop on Incentives and Trust in E-Communities, WIT-EC'15*, pp. 21–24, 2015.

[11] M. Nojoumian and D. R. Stinson, "Social secret sharing in coud computing using a new trust function," in *10th IEEE Annual International Conference on Privacy, Security and Trust (PST)*, pp. 161–167, 2012.

[12] Trust., *Definition of trust in English:*. oxford-dictionaries., 2016.

[13] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527–539, 1987.

[14] D. Power, *Types of Decision Support Systems*. gdrc.org, 2016.

[15] P. Madhavan and D. A. Wiegmann, "Similarities and differences between human–human and human–automation trust: an integrative review," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 277–301, 2007.

[16] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[17] B. M. Muir, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[18] B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[19] I. Dassonville, D. Jolly, and A. Desodt, "Trust between man and machine in a teleoperation system," *Reliability Engineering & System Safety*, vol. 53, no. 3, pp. 319–325, 1996.

[20] N. Moray and T. Inagaki, "Laboratory studies of trust between humans and machines in automated systems," *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 203–211, 1999.

[21] M. Itoh, G. Abe, and K. Tanaka, "Trust in and use of automation: their dependence on occurrence patterns of malfunctions," in *Systems, Man,*

TABLE IV. TRUST IN AUTOPILOT SYSTEMS

Reference	Summary	Approach	Concentration
[49]	Automatic failures have more negative impacts on trust than manual failures	Experiment using route planning simulation	Trust in automatic and manual failures
[50]	Provides a procedure for improving trust in air traffic management systems	Survey-based	Trust in air traffic management systems
[51]	Greater decrease in trust when systems continuously outputted false alarms	Experiments using bread boards	Trust in autonomous system and false alarms
[52]	Participants analyzed airplanes and rated them as friendly or hostile based on different factors	Simulation-based	Human trust and judgment in airplanes with autopilot features
[53]	Culture affects trust	Internet based survey	Trust in autonomous and semi-autonomous autopilot systems

and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on, vol. 3, pp. 715–720, IEEE, 1999.

[22] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[23] S. M. Merritt and D. R. Ilgen, "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*, vol. 50, no. 2, pp. 194–210, 2008.

[24] R. R. Murphy, D. Riddle, and E. Rasmussen, "Robot-assisted medical reachback: a survey of how medical personnel expect to interact with rescue robots," in *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pp. 301–306, IEEE, 2004.

[25] J. Carlson and R. R. Murphy, "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*, vol. 1, pp. 180–186, IEEE, 2004.

[26] R. Parasuraman and C. A. Miller, "Trust and etiquette in high-criticality automated systems," *Communications of the ACM*, vol. 47, no. 4, pp. 51–55, 2004.

[27] D. P. Stormont, "Analyzing human trust of autonomous systems in hazardous environments," in *Proc. of the Human Implications of Human-Robot Interaction workshop at AAAI*, pp. 27–32, 2008.

[28] article36, "ban-autonomous-armed-robots/," *article36*, 2012.

[29] D. Li, P. P. Rau, and Y. Li, "A cross-cultural study: Effect of robot appearance and task," *International Journal of Social Robotics*, vol. 2, no. 2, pp. 175–186, 2010.

[30] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.

[31] R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? the development of a human–robot interaction trust scale," *International Journal of Social Robotics*, vol. 4, no. 3, pp. 235–248, 2012.

[32] J. Penders, P. Jones, A. Ranasinghe, and T. Nanayakara, "Enhancing trust and confidence in human robot interaction," tech. rep., 2013.

[33] C. Castelfranchi and R. Falcone, "Trust and control: a dialectic link," *Applied Artificial Intelligence*, vol. 14, no. 8, pp. 799–823, 2000.

[34] Y. Wang, Z. Shi, C. Wang, and F. Zhang, "Human-robot mutual trust in (semi) autonomous underwater robots," in *Cooperative Robots and Sensor Networks 2014*, pp. 115–137, Springer, 2014.

[35] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-time sensing of trust in human-machine interactions," *IFAC-PapersOnLine*, vol. 49, no. 32, pp. 48–53, 2016.

[36] M. Abd, I. Gonzalez, M. Nojoumian, and E. Engeberg, "Impacts of robot assistant performance on human trust, satisfaction, and frustration," in *To appear in RSS: Morality and Social Trust in Autonomous Robots*, 2017.

[37] M. Abd, I. Gonzalez, M. Nojoumian, and E. Engeberg, "Trust, satisfaction and frustration measurements for real-time human-robot interaction," in *30th Florida Conference on Recent Advances in Robotics*, pp. 89–93, 2017.

[38] driverless future, "Forecasts," *http://www.driverless-future.com/6*, 2016.

[39] A. Uggirala, A. K. Gramopadhye, B. J. Melloy, and J. E. Toler, "Measurement of trust in complex and dynamic systems using a quantitative approach," *International Journal of Industrial Ergonomics*, vol. 34, no. 3, pp. 175–186, 2004.

[40] T. Helldin, G. Falkman, M. Riveiro, and S. Davidsson, "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*, pp. 210–217, ACM, 2013.

[41] D. Howard and D. Dai, "Public perceptions of self-driving cars: The case of berkeley, california," in *Transportation Research Board 93rd Annual Meeting*, 2014.

[42] M. S. Carlson, M. Desai, J. L. Drury, H. Kwak, and H. A. Yanco, "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*, pp. 20–27, 2014.

[43] M. Kyriakidis, R. Happee, and J. De Winter, "Public opinion on automated driving: results of an international questionnaire among 5000 respondents," *Transportation research part F: traffic psychology and behaviour*, vol. 32, pp. 127–140, 2015.

[44] M. Wagner and P. Koopman, "A philosophy for developing trust in self-driving cars," in *Road Vehicle Automation 2*, pp. 163–171, Springer, 2015.

[45] V. Butakov and P. Ioannou, "Driving autopilot with personalization feature for improved safety and comfort," in *18th International Conference on Intelligent Transportation Systems*, pp. 387–393, IEEE, 2015.

[46] W. Payre, J. Cestac, and P. Delhomme, "Fully automated driving impact of trust and practice on manual control recovery," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 58, no. 2, pp. 229–241, 2016.

[47] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *American Control Conference (ACC), 2017*, pp. 1542–1548, IEEE, 2017.

[48] R. A. Daziano, M. Sarrias, and B. Leard, "Are consumers willing to pay to let cars drive for them? analyzing response to autonomous vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 78, pp. 150 – 164, 2017.

[49] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 719–735, 2003.

[50] C. Kelly, "Guidelines for trust in future atm systems-principles," tech. rep., 2003.

[51] X. Jiang, M. T. Khasawneh, R. Master, S. R. Bowling, A. K. Gramopadhye, B. J. Melloy, and L. Grimes, "Measurement of human trust in a hybrid inspection system based on signal detection theory measures," *Int. J of Industrial Ergonomics*, vol. 34, no. 5, pp. 407–419, 2004.

[52] Y. Seong and A. M. Bisantz, "The impact of cognitive feedback on judgment performance and trust with decision aids," *Int. J of Industrial Ergonomics*, vol. 38, no. 7, pp. 608–625, 2008.

[53] S. R. Winter, S. Rice, R. Mehta, I. Cremer, K. M. Reid, T. G. Rosser, and J. C. Moore, "Indian and american consumer perceptions of cockpit configuration policy," *Journal of air transport management*, vol. 42, pp. 226–231, 2015.