Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

What is trust, actually? Review on the discussion regarding trust in technology and in AI

Saara Ala-Luopa
Human-centered AI for the Smart City
18.2.2022

Table of content

1 Introduction

Trust: we are all aware of its importance both in personal life and business. But defining trust is different. It is difficult. Because trust is a dynamic, complex, multidimensional, and underlying phenomenon. It escapes definitions and takes different forms according to the context and situation. There isn't one particular "trust". Despite the complex nature of trust, there is at least one common element in trust formation. To trust, there needs to be someone or something to trust on: *the other*. The trustee[1] can be another human, but it can also be a technological artifact. The trustor, in this case, refers to an individual human user.

This review aims to understand trust in technology and especially trust in artificial intelligence focusing on a human-centered perspective. We present trust in AI from three different perspectives: academic perspective, industry perspective, and governmental perspective. This paper aims to answer the following questions: a) What is trust in technology and in artificial intelligence, and b) how can trust in artificial intelligence be built and maintained? We will start from the beginning: psychology and sociology to understand the philosophical perspective on trust. Current research on trust in technology relies on these perspectives, therefore it is important to understand the antecedents of current technology trust research. Our aim is to keep this section relatively short: it will work as an introduction

[1] Trustor: entity that trusts, trustee: entity to be trusted

to the topic. After this we will present current academic research regarding trust in technology. This will be divided into different sections according to the thematic area: trust in technology in general, trust in automation, network of trust and trust in AI. Lastly, we will present industry and governmental perspective on trust in AI. In the discussion section, we will shortly discuss the differences between trust and trustworthiness and recap the review.

Trust in AI is widely discussed topic, but as a trust researcher this discourse does not aim to understand or define trust. Often trust seems to be just another buzzword in the current AI hype. With this review, we aim to concretize trust in AI *from the human-centered perspective* while honoring the existing research on trust in technology. However, as trust is dynamic by nature, it is linked to the current technological (and societal changes). Therefore, we also need to be able to approach this complex and intriguing phenomenon from holistic and human-centric perspective. As technology is embedded in society, also trust in technology becomes a societal phenomenon combining elements both from interpersonal trust and technological trust. This review provides an understanding of the mundane but undefined topic of trust, with a special focus of trust in novel, adaptive and autonomous AI technologies. The aim of this review is to make sense of the complex topic of trust.

Please note, that this is an ongoing work. This paper provides an overview of trust research and approaches focusing on human-centered perspective and aims to understand trust approaches from different perspectives. This work has been driven by the authors' personal research interest offering a background information for qualitative research regarding trust in AI in expert work, and experts' user experience.

2 Trust in society: academic background

2.1. Trust in social psychology

Social psychology emphasis trust as interpersonal and **individual personal characteristic**. Therefore, some people are more eager to trust other humans whereas some people are more distrusting. This is affected by, for instance, previous personal trust relationships and experiences. According to Julian Rotter, American psychologist and researcher:

> *"Trust is an expectancy held by an individual that the word, promise or written communication of another can be relied upon"* (Rotter, 1967). "

Rotter (1967) refers to trust as an **expectation** that he defines "generalized expectancy". This conception of trust follows a *social learning theory* approach. Social learning theory suggests that social behavior is learned by observing and imitating the behavior of others (Bandura, 1977). Rotter (1971) suggests that expectations for a particular situation are determined by specific previous experiences with situations that are perceived to be similar. According to Rotter (1967), the efficiency, adjustment, and even survival of any social group depends upon the presence or absence of trust.

American social psychologist John Rempel et al. (1985) perceive trust **dynamic** and evolving phenomenon, with the basis of trust changing as the relationship progressed. Similarly, to Rotter, they define trust as an expectation:

> *“Trust is an expectation related to subjective probability an individual assigns to the occurrence of some set of future events”* (Rempel et al. 1985).

Rempel et al. (1985) identify three coherent dimensions of trust that influence people's acceptance of the information provided by an external source: **predictability, dependability, and faith.** These components emphasize trust in interpersonal relationships and especially between romantic partners. *Predictability* forms the basis of trust early in a relationship, which is the degree to which future behavior can be anticipated. This is followed by *dependability*, which is the degree to which behavior is consistent. As the relationship matures, the basis of trust ultimately shifts to *faith*, which is a more general judgment that a person can be relied upon.

The dynamic characteristic of (interpersonal) trust has been utilised also in trust in automation. Bonnie M. Muir (1987) employed the trusting attributes found in Rempel et al. (1985). Over time, trust in automation evolves alongside these three dimensions: *predictability, dependability, and faith*. Reliability refers to the extent the technology responds to similar circumstances at different points in time (Muir, 1987; Muir and Moray, 1996). A similar progression emerged in Shoshana Zuboff's study of operators' adaptation to new technology (Zuboff, 1988). According to Zuboff, trust in this context depended on trial and error experience, followed by an understanding of the technology's operation, and finally, faith.

2.2. Trust in organization and management science

Possibly the best-known and the most widely accepted definition of trust emerged from organization and management science. In 1995 Roger C. Mayer, James H. Davis and F. David Schoorman published their article “An Integrative Model of Organizational Trust”. Their model draws from multiple perspectives and is stated to be applicable in multiple domains. Trust in this article is states as:

> *“Trust is the **willingness** of a party to be **vulnerable** to the actions of another party based on the **expectation** that the other will perform a particular action important to the trustor, irrespective of the **ability to monitor or control** that party.”* (Mayer et al. 1995)

According to Mayer et al. (1995) three attributes affect interpersonal trust: *ability, integrity and benevolence*. **Ability** is the group of skills, competencies and characteristics that enable the trustee to influence the domain. **Integrity** is the degree to which the trustee adheres to a set of principles the trustor finds acceptable. **Benevolence** is the extent to which the intents and motivations of the trustee are aligned with those of the trustor. They see trust as a *belief* that a person or technology has the attributes necessary to perform as expected in a situation.

Denise M. Rousseau et al. (1998) agree with this definition emphasizing the elements of **vulnerability, risk, and positive expectations** in trust formation. Trust is the assured reliance on the character, ability, strength or truthfulness of someone or something. Rousseau et al. (1998) define trust in the following way:

> *"Trust is a psychological state where an individual accepts vulnerability based upon positive expectations of the intentions or behavior of another – trust is the willingness to be vulnerable based upon positive expectations of the intentions or behavior of the other party."* (Rousseau et al. 1998)

2.3. Trust in sociology

Sociologists approach trust from a systemic perspective. Niklas Luhmann, a German sociologist, defines trust as a *system trust*: the system is expected to operate in **reliable** manner (Luhmann 1989). In addition, trust is an essential part of the society. Trust is the

> " ***social glue*** *that holds everything in society together"*- - and "*an operator's belief that a system is operating in a* ***predictable manner*** *and in keeping with* ***expectations****." (*Luhmann 1979, 1988, see also Lewis and Wiegert 1985).

Luhmann approaches trust from the interpersonal perspective: it is a way to reduce tensions and social complexity especially in a situation where the social environment cannot be regulated through rules and customs. Thus, people adopt trust as a central social complexity reduction strategy, and trust serves as a mechanism **to reduce perceived social complexity** (Luhmann 1979). According to Luhmann, trust refers to a behavior in which **familiarity** plays a part. For trust to exist, past experiences are needed to establish familiarity with the situation, for example, familiarity with a given technology:

" Trust is the willingness to behave based on expectation about the behavior of others when considering the risk involved; with an a priori trustworthy party, familiarity builds trust because it creates an appropriate context to interpret the behavior of the trusted party; trust is the product of fulfilled expectations." (Luhmann, 1979)

According to Polish sociologist Piotr Sztompa (1999), "*trust is a social construction that originates from interpersonal relationships.* Sztompka known for his work on the theory of **social trust**, where he assigned trust an additional feature. In this view, trust is more than just passive consideration of future possibilities: trust is a conviction-based approach, which means that only the actions taken when faced with **uncertainty** by the trusting party are the evidence of trust in the other party of the relationship. Sztompka states that the growing interest in the notion of trust is primarily due to the growing uncertainty surrounding the phenomenon and the **need for risk taking**, the growing interdependence and the need for cooperation, the growing number of new threats and dangers, and the unrestricted ability to make choices that increase the level of uncertainty. He has argued that trust in a person and trust in a technology are not fundamentally different, because behind all human-made technologies, there stand people who design, operate, and control them.

3 Trust in technology: academic perspective

3.1.1 Trust in technology in general

Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

Trust in technology has been studied both in IS and HCI fields, the latter focusing more on experimental research and the former on theorizing or conceptualizing the subject. It is noteworthy that trust in technology has been questioned because computers are not "moral agents" and computer do have the free will, and therefore the concepts of motivation and trustworthiness do not apply. (Friedman et al. 2000; Solomon & Flores, 2001). On the other hand, other researchers see computers as social actors to which people respond according on social relationship-like rules: technologies are social actors in the sense that they have a social presence, and people respond to this social presence (Corritore et al. 2003, Nass et al. 1996; 1995). Riegelsberger (2005) states that in many cases trust in technology will be linked to trust in the socio-technical systems which this technology is part of. When technology is embedded in society, it will include moral justification as well.

Previous research on trust in technology has widely accepted the trust definition from organization and management science by Mayer et al. (19095), where trust is defined as:

> *" the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party"* (Mayer, Davis, & Schoorman, 1995).

This definition includes the element of *expectation* regarding behavior or outcome. Trust is described in terms of a **performance of an agent** that furthers the **goals of an individual** who is dependent on the trustee. Trust concerns an ***expectancy*** or an attitude regarding the likelihood of favorable responses. Trust describes a relationship that depends on the characteristics of the trustee, the trustor, and the goal-related context of the interaction. In interpersonal relationships these characteristics are *ability, integrity, and benevolence* (Mayer et al., 1995). However, in trust in technology these human-like characteristics might not apply. Therefore, some researchers have explored system-like characteristics. For instance, Lankton & McKnight (2011) propose three technology-related trust beliefs: *reliability, functionality and helpfulness,* to parallel the interpersonal trust beliefs. **Functionality** is the degree to which one anticipates the technology will have the functions or features needed to accomplish one's tasks. **Reliability** is the degree to which an individual anticipates the technology will continually operate properly or will operate in a consistent flawless manner. **Helpfulness** refers to the degree to which an individual anticipates the technology will provide adequate and responsive help. Similar conceptualization was made by Lippert (2001), and Muir and Moray (1996).

However, if technology has human-like characteristic, this might create human-like trust. Reeves and Nass (1996) define **human-like technologies** as those that are inherently communicative, instructive, and interactive. They found that when people interacted with technology that provided representations of humans, people interacted with the technology as if it were a person. Similarly, Wang and Benbasat (2005) studied recommendation agents that have human-like characteristics: these agents ask questions that allow humans to respond, provide information to humans, and sometimes mimic person-to-person communication. According to this study, consumers treat online recommendation agents as "social actors" and perceive human characteristics (e.g., benevolence and integrity) in computerized agents. Also, Corritore et al. (2003) see technology as social actor due to their social presence: the

extent to which a medium allows users to perceive others as being physically present (Fulk et al. 1987). Technology have been shown to have more social presence if they exhibit a strong social richness, incorporate personalization and human images, or incorporate human audio or video (Cyr et al. 2009; Gefen and Straub 2004; Lombard and Ditton 1997). Therefore, how technical or social the technology is perceived, might influence trust formation.

Table 1 offers a great summary of the conceptual origins of technology trust beliefs (Lankton & McKnight 2011).

Technology Trust Beliefs	Conceptual Origins			
	Trust in Information Systems	Trust in Online Environments and Online Agents	Trust in Automation	Interpersonal Trust
Functionality The degree to which one anticipates the technology will have the functions or features needed to accomplish one's task(s)		*Competence* The trustee has the ability, skills, and expertise to perform effectively in specific domains (Wang and Benbasat, 2005, p. 76).	*Competence* The extent the technology performs its functions properly (Muir and Moray, 1996, p. 434)	*Ability (Competence)* The group of skills, competencies, and characteristics that enable a party to have influence with some specific domain (Mayer et al., 1995, p. 717).
Reliability The degree to which an individual anticipates the technology will continually operate properly, or will operate in a consistent flawless manner.	*Reliability* The technology is fully functioning and not experiencing system downtime when completing job related tasks (Lippert, 2001).	*Integrity* An individual believes that the trustee adheres to a set of principles (Wang and Benbasat, 2005, p. 76).	*Reliability* The extent the technology responds similarly to similar circumstances at different points in time (Muir and Moray, 1996, p. 434).	*Integrity* The trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable (Mayer et al., 1995, p. 719).
Helpfulness The degree to which an individual anticipates the technology will provide adequate and responsive help.		*Benevolence* The trustee cares about her and acts in her interests (Wang and Benbasat, 2005, p. 76).		*Benevolence* The extent to which the trustee is believed to want to do good to the trustor, aside from an egocentric motive Mayer et al., 1995, p. 718).

3.1.2 Trust in automation

Previous research on trust in automation offers useful background to study trust in autonomous and adaptive AI technology. Many studies have demonstrated that trust is a meaningful concept to describe human-automation interaction, both in naturalistic settings (Zuboff, 1988) and in laboratory settings (see e.g., Lee & Moray, 1992; Muir & Moray, 1996). According to Lee & See (2004), two critical elements define the basis of trust. The first is the focus of trust: what is to be trusted? The second is the type of information that describes the entity to be trusted: what is the information supporting trust? This information guides expectations regarding how well the entity can achieve the trustor’s goals.

Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

Trust in automation is again own concept and differs from regular technology trust. Commonly accepted definition of trust in automation is made by Lee & See (2004). In their thorough review regarding trust in automation, they define trust as:

> *“ The attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.”* (Lee & See, 2004).

This definition is used also in the research regarding trust in AI. Lee & See (2004) argue that this definition must be elaborated to consider the **appropriateness of trust, the influence of context, the goal-related characteristics of the agent, and the cognitive processes** that govern the development and erosion of trust.

Similarly, to Mayer et al. (1995) suggestion of ability, integrity, and benevolence as the bases of trust, Lee and Moray (1992) made similar suggestion to define the trust relationship between human operator and automatized system. They defined the following factors as the general bases of trust in automation: *performance, process, and purpose.*

Performance refers to the current and historical operation of the automation and includes characteristics as reliability, predictability and ability. Performance information describes what the automation does. Because performance is linked to ability to achieve specific goals, it demonstrates the task- and situation-dependent nature of trust. **Process** is the degree to which the algorithms of the automation are appropriate for the situation and able to achieve the operator’s goals. Process information describes how the automation operates and is similar to characteristics as dependability and integrity. Process focuses on specific behaviors, qualities and characteristics attributed to an agent: trust is in the agent and not in the specific actions of the agent. **Purpose** refers to the degree to which the automation is being used within the realm of the designers’ intent and describes why the automation was developed. Purpose corresponds to faith and benevolence and reflects the perception that the trustee has a positive orientation towards the trustor. According to Lippert (lähde), purpose refers to the expected benefit of the technology which is based on the organization culture and values.

Muir (1994) evoked a discussion of trust between people: an operator (user) simply cannot have complete knowledge of an automated system. The operator’s perceptions become important because of the automation’s **freedom to act**, and the trustor’s i**nability to account for all possibilities of the trustee’s action**. This is similar to trust in AI which might include black box technologies, autonomous and adaptive behavior, and non-intended consequences. Muir (1987) built upon the interpersonal trust attributes found in Rempel et al. (1985): predictability, dependability, and faith. In phases when knowledge is still low, trust is mainly driven by the perceived **predictability** of a technology. This is followed by **dependability** and **faith**. However, trust in automation can also follow an opposite pattern, where faith is important early in the interaction, followed by dependability, and then by predictability (Muir & Moray, 1996). Muir conceptualizes trust as an expectation: technology is expected to act in a reliable and consistent way in the future, which leads to the individual evaluation or assessment of the predictability of the technology.

Zuboff has studied trust in naturalistic settings in her book “*In the age of smart machines: The future of work technology and power”*. This book describes an ethnographic field study

of operators adapting to computerized manufacturing systems and illustrates how IT shape workers and work in organizations. Zuboff (1988) found that trust in a new technology depends on **trial-and-error experience**, followed by **understanding of the technology's operation, and finally, faith**. These results show that trust in automation, like trust in people, is culturally influenced in that it depends on long-term experiences of a group of people. Zuboffs' work has been used to explain current characteristics of IT, such as organizational dynamics and power differentials in organizations (Burton-Jones, 2014), as well as AI's dual capacities for automating and informating work (Jarrahi, 2019)

Hoff and Bashir (2015) synthesize empirical research on trust in automation. They argue, that for instance performance – purpose – process framework is applicable only to relationships with unfamiliar trustees and automated systems. However, in reality, the trust formation process depends on a number of factors related to the operator, environment, and automated system. In their review, they conceptualize that variability by systematically reviewing empirical research on trust in automation. Based on this review, they define three sources of variability in trust in automation: *dispositional, situational, and learned trust.* **Dispositional** factors include the age, culture, and personality of the trustor among other characteristics. **Situational** factors concern the context of the human-automation interaction and various aspects of the task, such as workload and risk. **Learned** trust is a result of system performance characteristics as well as design features that present how performance is interpreted. They argue that designers can better facilitate appropriate trust by providing users with ongoing feedback concerning the reliability of automation and the situational factors that affect its performance. In order to promote greater trust and discourage automation disuse, designers should consider increasing an automated system's degree of anthropomorphism, transparency, politeness, and ease-of-use.

Misuse and disuse are concepts that define the flawed partnership between automation and people. **Misuse** refers to the failures that occur when people inadvertently violate critical assumptions and rely on automation inappropriately, whereas **disuse** signifies failures that occur when people reject the capabilities of automation (R. Parasuraman & Riley, 1997). Supporting appropriate trust is critical in avoiding misuse and disuse of automation. Calibration refers to the correspondence between a person's trust in the automation and the automation's capabilities (Lee & Moray, 1994; Muir, 1987). The definitions of appropriate calibration of trust parallel those of misuse and disuse in describing appropriate reliance.

3.1.3. Network of trust

Interestingly, Muir (1994) has developed a concept of a *network of trust* to identify and analyse the different trust relationships in complex technical systems. A **network of trust** consists of the different parties prevalent and the trust relationships in which the parties are engaged. In their work, the parties are: 1) designers, 2) the system, 3) operator 1, 4) operator 2 (accounting for the fact that a system is run by multiple operators that share or trade tasks), 5) management and 6) society. According to the concept, these parties share different kind of trust relationships. For instance, designers, operators, and management share mutual trust relationships. Management needs to trust the operators to control the system correctly, while the operators are asked to trust the policy decision, for example, safety/productivity trade-

offs, made by management. System and society instead do not share mutual trust relationship with the other parties, but only take the role of a trustor (only giving trust, society) or a trustee (only receiving trust, the system). However, all parties involved need to trust the system to be useful in the particular context. Society, on the other hand, needs to trust all other parties involved to run the system safely.

The concept of *network of trust* is interesting and ahead of time because it aims to understand trust in technology from the wider sociotechnical perspective, rather than focusing solely on the technical solution. Söllner et al. (2012, 2013) have continued Muir's (1994) work in the context of IS use and among these relationships. They argue that this approach emphasizing multiple trust relationships should be used when studying trust in the context of IS use, since different trustees resemble different targets of trust, and thus are prevalent. In their work, they identify five relevant targets of trust from a user's point of view (=parties), which are: 1) the user, 2) IS, 3) Internet, 4) provider and 5) community of Internet users.

3.1.4 Trust in AI

Trust in AI is a hot topic today, both in academic research and in industry. Although there are different approaches to trust in AI (e.g., technical and societal), the common element is linked to previous research of trust in technology: trust is essential in the use and acceptance of novel technologies. There is no clear definition of trust in AI and due to AI's varied implementation, it might be difficult to create such definition. More fruitful approach is to explore, how trust in AI is approached, which perspectives emphasize and how previous research of trust is utilized in these studies. This can help to establish a conceptualization for human-centered trust in AI. These articles focus on empirical, qualitative research regarding trust in AI. *This is not a thorough review: the aim is to offer examples, how trust in AI can be studied.*

Hengstler et al. (2016) were among the first one to conduct empirical, qualitative research regarding trust in AI. They conducted interviews with industry experts who were scientists from engineering and traffic research institutes to explore, how firms systematically foster trust regarding applied AI. They analyze nine case studies in the transportation and medical technology industries following an inductive, multiple case study research (Eisenhardt, 1989). This research is based on the work of Lee and Moray (1992), who identified **performance, process, and purpose** as the general bases of trust. Their findings reveal that trust in applied AI requires not only trust in the technology but also trust in the innovating firm and its communication. They argue that trust in applied AI is an evolving and dynamic phenomenon. Consequently, firms must begin to build trust during the democratic development process of an applied AI. The performance basis is primarily reliant on both **operational and data security** aspects, the process basis is determined by **cognitive compatibility, trialability, and usability**, and the purpose basis is founded on **application context and design.** Trust in the innovating firm increases with stakeholder alignment, high public transparency of the development project, and gradual introduction of the technology. Ultimately, trust in communication grows primarily by early, proactive, and application-based communication as well as the transmission of benefit-related information. This study

Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

suggests that trust in AI is not based only on technical characteristics but is formatted also in human-human communication and collaboration.

Lee (2018) studied trust in AI from the perspective of algorithmic decision-making. They conducted a between-subjects online experiment using a scenario-based method: participants read descriptions of managerial decision that either algorithms or people had made. The goal of this study was to understand the perceptions of algorithmic decisions in management contexts, and how perceptions differ depending on whether the decision-maker is a person or a machine. The managerial decisions were based on real-world examples of workplaces where algorithms have begun to change organizational practices. Then they examined the influence of the decision-maker (algorithmic or human) on participants' perceptions of the decisions. Their study is based on the social psychology of computing technologies, and on emerging theories around people's experiences with algorithmic technologies. To define trust, they use **Lee & See (2004) definition** as trust as "*an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability* ".

These results suggest that what people think algorithms are capable of and their comparison with human decision-makers play important roles in people's judgments of trustworthiness and fairness, as well as their emotional responses – regardless of the actual performance of algorithms. Their results suggest that **task characteristic**s, particularly whether the task is better suited to human or mechanical skills, significantly influence how people perceive algorithmic decisions compared to human-made ones. With the mechanical tasks, algorithmic and human-made decisions were perceived as equally fair and trustworthy. However, human managers' fairness and trustworthiness were attributed to the manager's authority, whereas algorithms' fairness and trustworthiness were attributed to their perceived efficiency and objectivity. Human decisions evoked some positive emotion due to the possibility of social recognition, whereas algorithmic decisions generated a more mixed response – algorithms were seen as helpful tools but also possible tracking mechanisms. With the human tasks, algorithmic decisions were perceived as less fair and trustworthy and evoked more negative emotion than human decisions. **A**lgorithms' perceived lack of intuition and subjective judgment capabilities contributed to the lower fairness and trustworthiness judgments. Positive emotion from human decisions was attributed to social recognition, while negative emotion from algorithmic decisions was attributed to the dehumanizing. Logg (2017) has studied how decision-makers expertise, the potential to affect oneself or others, and the subjectivity and objectivity of decisions can influence people's reliance on algorithmic advice.

It is noteworthy, that how we use technology reflects our values: what decision or work tasks we automate and what we keep "human", are decisions. What is the profession we might underestimate (as machine can do it similarly as human can); or what are the professions we believe could be enhanced with AI technologies (as human are not doing it well enough?) How we perceive certain professions and value the humanness, education, or the holistic perspective are reflected in the technology development. This also reflects the comparison between human and the technology: which party is more reliable in certain context and why. Maybe in the future we have the possibility to decide, if we want AI-decision or humanmade-decision?

Ashoori et al. (2019) studied factors that influence trustworthiness of AI-infused decision-making processes. More specifically, they aim to understand how different factors about a

decision-making process, and an AI model, influences peoples' perceptions of the trustworthiness of the process. They use a scenario-based approach in which a decision-making process is described, and let the participants rate their attitudes toward it. Factors are, e.g., decision stakes, decision authority, and model interpretability. Their evaluation of trust focused on several, the most relevant dimensions: overall trustworthiness, reliability, technical competence, understandability, and personal attachment. Similarly, to Lee (2018), they **adopt definitions of trust proposed by Lee and See** (2004).

Their findings suggest that designing trustworthy AI is a difficult task and navigating it successfully will require deep input from not only the people building the AI, but from the people using it and from the people affected by it as well. There is no "magic formula" that can be prescribed to guarantee that an AI system is trusted. They encourage for additional work to understand the abilities of different types of **explanations** in building trust in an AI system and underline a need for **multistakeholder perspective.** More research is needed to understand what information is required to establish trust amongst the different stakeholders in an **AI system's lifecycle:** the people building the system, the people using the system, and the people affected by it. They urge researchers and AI system designers to also consider the mistrust and distrust in AI design and development, and how to design these elements. Similar emphasis was also visible in industry reports, e.g., Google.

Zhang et al. (2020) **c**onducted a case study of AI-assisted decision-making and examine the impact of information designs that reveal case-specific model information, including confidence score and local explanation, on people's trust in the AI and the decision outcome. In AI-assisted decision making the individual strengths of the human and the AI come together to optimize the joint decision outcome. A key to the success is to appropriately **calibrate human trust** in the AI on a case-by-case basis; knowing when to trust or distrust the AI allows the human expert to appropriately apply their knowledge, improving decision outcomes in cases where the model is likely to perform poorly.

The findings show that confidence score can help calibrate people's trust in an AI model. However, trust calibration alone is not sufficient to improve AI-assisted decision making, which may also depend on whether the human can bring in enough unique knowledge to complement the AI's errors. This study is interesting, because it offers an empirical example of **human-AI collaboration** and emphasizes the **critical attitude and contextual information** in this trust relationship. User must be able to assess technology and then device, whether to follow AI provides outcome or not.

Glikson & Woolley (2020) conducted a review with an aim to explain how AI differs from other technologies. They present the existing empirical research on the determinants of human "trust" in AI, and identify the form of AI representation (robot, virtual, and embedded) and its level of machine intelligence (i.e., its capabilities) as important antecedents to the development of trust. Based on this, they propose a framework that addresses the elements that shape users' cognitive and emotional trust. For each AI representation (robotic, virtual, and embedded), they discuss the common dimensions that emerged from the review as relevant for cognitive trust (tangibility, transparency, reliability, task characteristics, and immediacy behaviors) and for emotional trust (tangibility, anthropomorphism, and immediacy behaviors). Cognitive trust is based on perceptions of trustee **reliance and competence.** When researchers examine cognitive trust in AI, they measure it as a function of **whether users are willing to follow information or advice** and

act on it, as well as whether they see the technology as helpful, competent, or useful. In this study, trust is defined as "*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (Mayer, Davis, & Schoorman, 1995).

The results of this review reveal the important role of AI's tangibility, transparency, reliability, task characteristics and immediacy behaviors in developing cognitive trust, and the role of AI's anthropomorphism specifically for emotional trust. **Tangibility** refers to AI's capability of being perceived or touched in developing trust (emphasizes in trust in virtual AI). **Transparency** is the level to which the underlying operating rules and inner logics of the technology are apparent to the users. **Reliability** means exhibiting the same and expected behavior over time. In the case of AI, reliability is often difficult to assess, especially in the context of high machine intelligence, as learning from data can lead technology to exhibit different behavior, even if the underlying objective function remains the same. **Task characteristics** refers to the work the technology is performing, such as whether it deals with largely technical or interpersonal judgments. **Immediacy behaviors** refer to the special interactive abilities of AI or socially oriented gestures intended to increase interpersonal closeness, such as proactivity and responsiveness. Especially transparency, (low) reliability, task characteristics (human task vs. AI task), and immediacy behavior (as personalization) increase trust in embedded AI.

3.2. Industry perspective

Industry approach to trust in AI depends regarding the company: other companies underline the technical perspective in trust formation, while others adopt more holistic and societally oriented approach. **Explainability**, and designing explainable AI is a common theme and essential element in building trust in AI. For instance, Google argues that "e*xplaining predictions, recommendations, and other AI output to users is critical for building trust."* Explainable AI creates **transparency**: why and how AI will end up into certain outcomes, recommendation, decision, prediction, or action (Ericsson, SiloAI, Google).

Google has created an entire website for **explainable AI,** offering practical guidelines to build explainable AI. They introduce in a detailed manner how user trust can be established in different **interaction points.** Google defines trust as dynamic, introducing ways to build trust in different points, e.g., before and during the interaction. Interestingly, they also underline aspect of **distrust or trust calibration.** They argue that users should not overtrust technology. AI should be able to inform the user of possible faults or unpredictability, as well as critical "trust moments" to create the appropriate trust. They have a very human-centric approach to the topic: explainability is seen to increase understanding and therefore trust. Google refers to Mayer et al. (1995) study and emphasizes ability-benevolence-reliabity factors that affect users' trust formation.

According to Capegmini, **transparency** helps to evaluate the technology: what AI does, how AI works and how it will be used and why. They underline the developers' understanding of the technology. Unless developers understand how the technology works, it cannot be trusted. First, AI should be transparent, then it can be designed to be either interpretitive or explainable, and therefore, understandable. Capegmini presents trusted AI

framework which combines technical trust and ethics of technology. They define the **development of trusted AI in three phase**s: a) discovery phase (partnership, team, data representativeness), b) training phase (documentation, training, explainability, c) deployment phase (model trustworthiness, responsibility areas. Capegmini emphasizes why-question and underlines human-centric perspective in AI design and development: Explainability, fairness, openness, and all those factors of trust in AI can only be fully answered when we know the “why” of the system and its goals.

It is noteworthy that explainability might not automatically create trust because it can confuse the user and distort the user experience. For instance, Ericsson underlines the meaning of **different user groups** and required information according to each stakeholder. Similarly, Deloitte has designed different **trust strategies for different user groups**. Their aim is to prevent possible **risks** by creating understanding to the users: trust is dynamic, and risk is essential element in trust formation. To trust, user needs to identify and understand the possible risks. Thus, they can control the technology.

IBM emphasizes holistic, even ethical, perspective in trust in AI. They underline **moral, transparency, education and collaboration** in the design and development of AI applications. They argue that **understanding** of the technology is essential in AI trust and this is intertwined with transparency. Transparency refers to developers: they must be transparent of the user **data collection**, for instance, what kind of data is collected, how data is used, and what kind of control the users have over their data. IBM sees trade-off between utility and privacy, but the user must be able to make decision regarding this trade-off. IBM emphasized AI education, and this reflects also the potential impact AI has on work-life and work practices, for instance, what comes to new required skills: how AI might affect the work-life and what kind of changes this could bring. They discuss about values and how to design values and ethics to AI applications. IBM **questions users’ trust** in technology: “w*hat level of trust can—and should—we place in these AI systems?”* On the other hand, they question how user should perceive AI: AI should not be compared to humans but perceived as another entity. The most important thing in trust is to understand how AI works and why it does certain decisions. IBM argues that trustworthy AI is human-centered, and underlines needs, safety and privacy. AI should also respect our data and be both transparent and explainable.

Accenture defines **Responsible AI** as the practice of designing, building and deploying AI in a manner that it empowers employees and businesses and fairly impacts customers and society. Also, Microsoft underlines responsible AI and responsive management and organizational perspective in developing and designing trustworthy AI: they emphasize **inclusive technology and security, privacy, and safety** as building blocks of trust in AI. Accenture criticizes that many organizations begin their journey by focusing on one issue, such as algorithmic fairness or compliance. However, most successful companies understand the importance of investing in all four pillars of **responsible AI**. According to Accenture, responsible AI implementations are: 1) Organizational: Democratic way of working and facilitation of human + machine collaboration, 2) Operational: Setting up governance and systems for AI 3) Technical: Ensuring systems and platforms are trustworthy and explainable by design, 4) Reputational: Articulating the Responsible AI mission and ensuring it is anchored to the company’s values, ethical guardrails, and accountability

structure. Responsible AI is, according to Accenture, organizational process. Trust inside and outside of the organization is a key component to getting to value from AI. Responsible AI principles must be put into practice. Similarly, to Deloitte, Accenture underlines **risk** as the biggest factor in trust formation. They have conducted a global survey, where 58% of risk managers fear AI causing unintended consequences over the next two years. Only 11% describe themselves as fully capable of assessing the risks in AI adoption in their organization.

KPGM studied citizens perceptions in AI in Australia. According to their findings, the public questions whether AI is designed to operate with **integrity and humanity**. This refers to interpersonal trust, and for instance, benevolence factor in trust formation. They argue that **lack of confidence in commercial organizations** to develop and regulate AI might be caused by people’s perception of the developers' purpose: they might underline financial motivation (to cut labour costs and increase revenue) rather than societal motivation (to help solve societal problems and enhance societal wellbeing). Their findings reveal that the public have very clear expectations of the principles and practices they expect AI systems to uphold in order to be trusted. Organizations are expected to maintain high standards of AI systems in terms: a) performance and accuracy, b) data privacy, security and governance, c) transparency and explainability, d) accountability, e) risk and impact mitigation, f) fairness g) human oversight. These principles and practices reflect to recent government reports on trustworthy AI. According to their survey, trust is influenced by four key drivers: 1) beliefs about the adequacy of current regulations and laws to make AI use **safe**, 2) the perceived uncertain **impact** of AI on society, 3) the perceived impact of AI on jobs, 4) familiarity and **understanding** of AI. They argue, that without public confidence that AI is being developed and used in an ethical and trustworthy manner, it will not be trusted, and its full potential will not be realized.

To summarize, the industry perspective in AI trust formation is not very coherent. One reason might be the difference in their business practices: some companies are developing companies while other might mainly focus on consultant. Overall, trust is important and key element in the use and acceptance of novel technologies. **Understandability, explainability and responsibility** are intertwined with trust formation. Trust in AI is not targeted only to technology or created in between user and technology only: developers or developing companies can also take a position of a trustee. Also, user has responsibility in understanding technology. Interestingly, both Google and IBM underline the element of **distrust** – users overtrust in AI is seen risky. Maybe even more important than trust, is the understanding of the AI. And this covers **different stakeholders with different needs for information.** It is easy to state, that trust is dynamic in horizontal way (through the interaction), but trust is also vertical as trust in different levels can be identified: technical, social and process levels. In all these stages, trust can, and should, be built in different way. Trust in AI is very holistic by nature.

3.3 Governmental perspective

The *European Commission* has published a European AI strategy (2018) which emphasizes human-centric AI. The argue that trust is a prerequisite to ensure a human-centric approach to

AI: AI is a tool that has to serve people with the ultimate aim of increasing human well-being. In order to achieve **‘trustworthy AI’**, three components are necessary, and these should be met throughout the system's entire life cycle: 1. AI should be **lawful**, complying with all applicable laws and regulations; 2. AI should be **ethical**, ensuring adherence to ethical principles and values; and 3. AI should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Ideally, all three work in harmony and overlap in their operation. In practice, however, there may be tensions between these elements. Guidelines emphasize the individual and collective responsibility in society to work towards ensuring that all three components help to secure Trustworthy AI.

Addressed to all stakeholders, guidelines seek to go beyond a list of ethical principles, by providing guidance on how such principles can be operationalized in socio- technical systems. Both technical and non-technical methods can be used for their implementation. **Seven key requirements for Trustworthy AI:** (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

Trust definition is based on literature, and in the guidelines, it is defined in the following way:*“Trust is a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements.”*

Their guidelines **emphasize trust between people**: in addition, that AI systems are legally compliant, ethically adherent and robust, trust can be ascribed to all people and processes involved in the AI system’s life cycle.

The *OECD Principles on Artificial Intelligence* (2019) promote AI that is innovative and trustworthy and that respects human rights and democratic values. These principles are the first signed up to by governments. Trustworthy AI refers to AI systems that embody the OECD AI Principles; that is, AI systems that respect i**nclusive growth, sustainable development and wellbeing; human rights and privacy**; are **fair, transparent, explainable, robust, secure and safe;** and the actors involved in their development and use remain **accountable**. The OECD AI Principles contain five recommendations for national policies and international co-operation. The recommendations include: 1) investing in AI research and development; 2) fostering a digital ecosystem for AI; 3) shaping an enabling policy environment for AI; 4) building human capacity and preparing for labour market transformation; and 5) international co-operation for trustworthy AI. Interestingly, OECD principles on AI present three types of tools that can be leveraged to facilitate implementation of the AI Principles. These tools can be classified as **technical, procedural, or educational.** This idea follows my personal perspective on different stages in trusting AI.

Technical tools for trustworthy AI aim to address specific AI-related issues from a technical angle, including bias detection, transparency and explainability of AI systems, performance, robustness, safety and security against adversarial attacks. They include toolkits, software tools, technical documentation, certification and standards, product development or lifecycle tools, and technical validation tools. A sizeable proportion of the technical tools submitted originate from large private sector companies, such as IBM, Google and Microsoft. Many of these technical tools to develop and use trustworthy AI exist as open-source resources, which facilitates their adoption and allows for crowdsourcing solutions to

software bugs. Many of these tools allow developers and others to check AI systems for **reliability** and fairness. **Objects: fairness, transparency, explainability, robustness.**

Procedural tools for trustworthy AI provide operational or process-related implementation guidance. They encompass guidelines, governance frameworks, product development, lifecycle and risk management tools, sector-specific codes of conduct and collective agreements, and process certifications and standards. Compared to technical tools, where there is high private sector participation, procedural tools to implement AI systems ethically and inclusively are **produced by a wider variety of stakeholders**, including **governments** and trade unions. Some procedural tools for transparency and explainability emphasize the importance of documenting the development and deployment of AI systems and propose governance frameworks for their implementation. **Objectives: Inclusive implementation, ethical implementation, transparent and explainable implementation.**

Educational tools for trustworthy AI encompass mechanisms to build awareness, inform, prepare or upskill stakeholders involved in or affected by the implementation of an AI system. They include **change management processes, capacity and awareness building tools, guidance for inclusive AI system design, and training programs and educational materials.** Depending on the implementation context, educational tools are designed to serve different audiences. They can be wide-reaching and open to the public at large or focus on a specific group affected by the implementation of an AI system, such as SMEs or workers. **Target audiences: Businesses, workplace actors, general public.**

National Security Comission on AI: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

3.4. Trust and trustworthiness of AI

The current discussion around AI focuses both on trust and trustworthiness / trustworthy AI. Sometimes these concepts might be a bit mixed. We perceive trustworthiness as the system characteristics and the AI design process. AI must be designed in a trustworthy manner: the role of systems designers and researchers is not solely to increase the functionality and usability of the systems but to design them to support trustworthy action and well-placed trust. According to Riegelsberger (2005), designers must be aware of their role as social engineers: AI systems will shape how people behave, and it will impact the level of trust and trustworthiness. Therefore, *trust* refers to the attitude of the trustor (“I trust X...”) whereas *trustworthiness* is the property of the trustee (“I trust because X is trustworthy”). I can trust in AI if it presents itself worthy of my trust.

In academic research, trustworthiness is defined in the following way:

- AI model is trustworthy to some contract if it can carry out its contract (Jacovi et al. 2020).
- Trustworthiness defines how trustworthy an actor is (Roy et al. 2016).
- The concept of “trustworthiness” refers to a multifaceted property of an aggregated source (Liu 2013).

Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

- AI is perceived as trustworthy by its users when it is developed, deployed, and used to adherence to general ethical principles (Independent High-Level Expert Group on Artificial Intelligence 2019).
- Trustworthy AI is related to normative statements on the qualities of the technology and typically necessitates ethical approaches (Toreini et al. 2020)
- Overall trustworthiness refers to the process ought to be trusted (Ashoori & Weisz)
- Trustworthy AI is human-centered and aims to offer high levels of human control, with the goal to lead to wider adoption and increase human performance, while supporting human self-efficacy, mastery, creativity, and responsibility (Razmerita et al. 2021).
- Trustworthy AI aims to contribute to the well-being of individuals as well as the prosperity and advancement of organizations and societies (Thiebes et al., 2021).
- Public expectations go beyond trusted systems: they want trustworthy systems based on respected independent oversight structures including professional organizations that develop effective voluntary guidelines and standards and government agencies etc. (Shneiderman, 2000, 2016).
- Trustworthy AI should respect all applicable laws and regulations, as well as a series of requirements: specific assessment lists aim to help verify the application of each of the key requirements (Floridi 2019)

Currently, there has been published few frameworks for trustworthy AI. For instance, Thiebes et al. (2020) introduce five foundational principles for trustworthy AI: (1) beneficence, (2) non-maleficence, (3) autonomy, (4) justice, and (5) explicability. Based on these principles, they develop a data-driven research framework for TAI. Their study is based on eight frameworks and guidelines developed and published by researchers, industry, and policymakers to promote (ethical) principles for trustworthy AI. The concept of trustworthy AI (TAI) promotes the idea that individuals, organizations, and societies will only ever be able to achieve the full potential of AI if trust can be established in its development, deployment, and use (adapted from EU: Independent High-Level Expert Group on Artificial Intelligence, 2019).

Zicari et al. (2021) outline a novel AI inspection process ‘Z-inspection’ based on applied ethics. They propose that Z-Inspection can be useful for auditing an AI system and to assess trustworthy AI in practice. The process can be used before production of an AI system, helping relevant actors to be aware of the ethical, social, technical, and legal risks and pitfalls when implementing an AI system. In defining trustworthy AI, they use the definition by the high-level European Commission’s expert group on AI: 1) human agency and oversight; 2) technical robustness and safety; 3) privacy and data governance; 4) transparency; 5) diversity, nondiscrimination, and fairness; 6) societal and environmental well-being; and 7) accountability. The Z-Inspection process is composed of three main phases:1) the Set Up phase, 2) the Assess phase and 3) the Resolve phase. They define **AI as an ecosystem** as a set of sectors and parts of society, level of social organization, and stakeholders within a political and economic context.

It is noteworthy, however, that trustworthiness does not guarantee trust. Sometimes we trust to those who are not trustworthy and vice versa. Sutrop (2019) states that establishing and articulating the purpose of trustworthy AI is not enough if people should trust AI systems: we also need to think about how to build trust in AI. Similarly, Vereschak (2021) states that

perceived trustworthiness is not recommended to be used as a proxy for how much there is trust. For instance, if the patient thinks that the doctor is trustworthy (e.g., has many diplomas, was recommended by someone), this does not mean the patient will trust them. To develop and design trustworthy AI, we must be able to create or design trustworthiness first. But we are not able to design trust as such – trust will be formatted in the evolving interaction between trustor and trustee. We can just aim to understand and affect the trust formation in the first place.

4 Discussion

Trust – a social glue, a key element in the use and acceptance of novel technologies, backbone for our relationships and an intriguing research topic. As seen above, trust can be approached from many disciplines and perspectives. It can build on, for instance, personal characteristics, situation, cultural context, technological attributes, previous experiences, and the overall process. To this end, studying trust requires strict definitions and justification based on an overall understanding of the topic. My aim is to study trust in AI from human-centered perspective: technology design and development emphasizes human needs and purpose of the technology (‘why instead of what’), what is the motivation behind technology development, and how it will affect humanity. To understand trust and to create appropriate trust, we need to approach trust from a holistic perspective. I believe that trust in AI, especially when embedded into society, combines both interpersonal trust and technical trust. Therefore, we need to understand societal and social structures that create trust in each point of the interaction or AI development process, in addition to the human-computer interaction. I believe that we are not able to create trust in AI if we focus solely on technical characteristics. To trust AI, we need to understand it, and we need to identify and overcome the possible risks in AI use. However, it is not as simple it seems. Firstly, technologies need to be transparent and explainable, but this explainability needs to be translated into the “common” language: for different stakeholders with different technical knowledge. Multi-stakeholder perspective emphasizes in human-centered trust in AI.

I agree with the notion, that trust is dynamic: it is a changing, evolving phenomenon which is most fragile in the beginning of the relationship. I believe, that in the beginning of trust in AI, the **personal aspect of trust** emphasize. This aspect includes society, media, organization, references and examples, personal interest – the inspiration, need or motivation to use certain technology that affects the users’ attitude. Social aspect helps the user to make an informed decision regarding the next phase. The second aspect is **social context:** understanding of the technology, risk identification, task characteristics, and goal setting, vendor or developers’ social presence or brand, and the ethical perspectives. The third aspect is **technical**, interaction with the product: trial and error experience, evaluation or assessment, transparency, explainability, reliability; human-computer interaction, leading to the use and acceptance of the technology. Even though I see trust as dynamic, these aspects are not completely linear. Instead, these elements remain in the background during the whole trust relationship, building in each other and supporting each other.

To this end, I present a definition of human-centered trust in AI:

Vipuvoimaa EU:lta 2014–2020
Euroopan unioni
Euroopan aluekehitysrahasto

> *“Human-centered trust in AI is holistic and includes personal, social, and technical elements. It is a dynamic and collaborative process and underlines purposeful technology use and a thorough understanding of the technology. Understanding towards AI will be created through transparent and explainable technology and communication, and it covers multiple stakeholders in different interaction points. At best, human-centric trust in AI manifests in an appropriate trust and responsible AI and maintain human control and autonomy while fulfills the potential in novel AI technology.”*

Please note, that this is the first conceptualization of “human-centric trust in AI” and requires more detailed work. This concept focuses on AI technology that is designed or developed with the vendor (tailored product/solution). In this case, the user can make an informed decision to use AI product. The AI aspect is not always clear because it might embed in the products (e.g., algorithms). Human-centric AI underlines that the user should be aware of AI in the products.

Another interesting, and essential, aspect of trust in AI is **appropriate trust** or trust calibration as **users’ critical attitude** towards the AI solutions. This is a novel perspective visible both in industry (e.g., Google) and in academia (e.g., Ashoori). According to this approach, designers should not build trust in the product too much because users should be able to calibrate the trust – they should be able to trust “correctly” and appropriately. This is explained by the possible unintended consequences AI might arouse due to its autonomous and adaptive nature. Rather, AI should inform the user of possible faults or unpredictability as well as critical “trust moments”. This perspective is one of my research findings as well, and I argue that it reflects a **paradigm shift in user experience regarding AI (AI UX)**: user should have the capability and tools to assess technology with critical attitude. Currently, we are used to technology that reduces our cognitive load and makes our lives easier. I do not see this current pattern to be a sustainable in the context of autonomous and adaptive AI technology, especially in work life and in professional domain.

I see **responsible AI** to closely intertwine with trust in AI. When we trust AI appropriately, we can use AI in responsible manner. Responsible AI is a collaborative goal. It is not task given only to technology developers. AI can’t have the moral responsibility – therefore, the developer, the users, the organizations and the society must have it. I believe that we have the capability to assess AI from the moral and ethical perspectives, if we understand it enough.

It is noteworthy that in organizational context, the individual user might not have the possibility to decide regarding their AI use. There might be differences how users perceive trust in technology in personal and in professional contexts. Also, there might be differences how lay users and expert users justify, evaluate and accept novel technologies, and what kind of new skills novel technologies might require from domain expert users. This depends on the scale and impact AI might have on different professional domains. *Expert user UX is a little studied subject and requires more detailed research.*

5 Conclusion

The rapid development of AI technology has sparked the discussion around trust in technology both inside and outside academia: also, industry and governmental actors are

participating the discussions. The importance of trust is understood and emphasized, yet many outcomes focus on general guidelines and requirements rather than a thorough understanding of trust formation. Common agreement seems to be that trust is essential in the acceptance and use of technology, and the autonomous, adaptive and societal nature of AI stresses trust.

I argue that AI is changing the relation between society and technology, and supporting this change requires a human-centric perspective on trust formation. Trust in AI is more than trust in human-computer interaction: it is dynamic, phased and holistic phenomenon which includes different stages. Social, collaborative and technical stages include different elements affecting trust formation. This conceptualization refers to human-centric AI and requires more detailed work and empirical research. In my PhD research, I will study this topic further.

References

Rotter, J. (1967). A New Scale for the Measurement of Interpersonal Trust. Journal of Personality, 35, 651-665.
http://dx.doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rotter 1971

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice Hall.

Rempel, J.K., Holmes, J.G. & Zanna, M.P. (1985). Trust in close relationships., 95-112. Journal of Personality and Social Psychology, 49.

Zuboff's 1988 book, In the Age of the Smart Machine: The Future of Work and Power, is a study of information technology in the workplace

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. The Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

Rousseau, Denise & Sitkin, Sim & Burt, Ronald & Camerer, Colin. (1998). Not So Different After All: A Cross-discipline View of Trust. Academy of Management Review. 23. 10.5465/AMR.1998.926617.

Luhmann, Niklas (1989) Ecological Communication. Cambridge: Polity Press. 218
Luhmann, Niklas (1989b) Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft, vol. 3. Frankfurt am Main: Suhrkamp.

https://helda.helsinki.fi/bitstream/handle/10138/23348/trustasa.pdf?sequen

Batya Friedman, Peter H Kahn, and Daniel C. Howe. 2016. trust Online. December 2000.

Cynthia L. Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: Concepts, evolving themes, a model. International Journal of Human Computer Studies 58, 6: 737–758.

Clifford Nass, B. J. Fogg, and Youngme Moon. 1996. Can computers be teammates? International Journal of Human Computer Studies 45, 6: 669–678.

Lankton & McKnight (2011)

Lippert (2001), and Muir and Moray (1996).

B. Muir, "Trust in Automation: Part 1. Theoretical Issues in the study of trust and human intervention in automated systems," Ergonomics, vol. 37, no. 11, pp. 1905- 1922, 1994.

Zuboff, S., 1988. In the age of the Smart Machine: The Future ofWork and Power. Basic Books, New York, NY.

Jarrahi, M. H. (2019). In the age of the smart artificial intelligence: AI’s dual capacities for automating and informating work. Business Information Review, 36(4), 178–187. https://doi.org/10.1177/0266382119883999

Burton-Jones, A. 2014. What have we learned from the Smart Machine? Information and Organization, 24, 71-105

Corritore, C.L., Kracher, B., and Wiedenbeck, S. 2003. "Editorial," International Journal of Human- Computer Studies (58:6), pp. 633-635

Vereschak, O., Bailly, G., Caramiaux, B., & How, B. C. (2021). ACM Reference Format: Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. Proc. ACM Hum.-Comput. Interact, 5, 39.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust-The case of autonomous vehicles and medical assistance devices. Technological Forecasting and Social Change, 105, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data and Society, 5(1). https://doi.org/10.1177/2053951718756684

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Logg, J. (2018). Theory of Machine: When Do People Rely on Algorithms? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2941774

Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. Retrieved from http://arxiv.org/abs/1912.02675

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295–305). Association for Computing Machinery, Inc. https://doi.org/10.1145/3351095.3372852

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Mayer, R. C., Davis, J. H. and Schoorman, F. D. “An Integrative Model of Organizational Trust,” Academy of Management Review (20:3), July 1995, pp. 709-734.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In Conference on Fairness, Accountability, and Trans- parency (FAccT ’21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445923

Toreini et al. (2020). The relationship between trust in AI and trustworthy machine learning technologies. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and TransparencyJanuary 2020 Pages 272–283. https://doi.org/10.1145/3351095.3372834

Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. Retrieved from http://arxiv.org/abs/1912.02675

Thiebes, S., Lins, S. & Sunyaev, A. Trustworthy artificial intelligence. Electron Markets 31, 447–464 (2021). https://doi.org/10.1007/s12525-020-00441-4