Building Trust in Artificial Conversational Agents

Clara Bove[a,b], Jonathan Aigrain[b] and Marcin Detyniecki[b,c]

[a]Sorbonne Université, CNRS, LIP6, Paris, France
[b]AXA, Paris France
[d]Polish Academy of Science, Warsaw, Poland

Abstract

The notion of trust on human and its construction mechanisms have been widely studies in the two last decades. However, research on this topic has focused on human interactions, on organizations or on some specific technologies such as the Internet or automation. With the increasing popularity of Artificial Conversational Agents (ACA), there is a growing interested to study trust in AI. In this paper, we present a new model of building trust in human-ACA interactions. Our contribution aims at proposing a theoretical heuristic for building trust in ACAs by combining multi-disciplinary research on human-to-human trust, trust in organizations and trust in technologies. We claim that an ACA is at the core of these three dimensions because it can be perceived as a technology by its nature, as a service provided to an individual with specific processes by its role, and as a social agent with its own set of intentions due to its human-like interaction model. We believe that this diversity in trust drivers is key to have a holistic approach of trust in ACAs.

Keywords

Artificial Intelligence; Trust; Human Computer Interaction; Artificial Conversational Agent; Design Guidelines; Interaction design.

1. Introduction

Artificial Conversational Agents (ACA) are now used in a wide range of applications, from medicine[1] to vocal assistants such as Amazon's Alexa or Apple's Siri. Given their popularity, it is now essential that ACAs are able to build a relationship of trust with users.

Interpersonal trust can be 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 part"[2]. However, we believe that the challenge of building trust in ACAs cannot be tackled through interpersonal trust only [3]. An ACA can be perceived as a technology by its nature, as a service provided to an individual with specific processes by its role, and as a social agent with its own set of intentions due to its human-like interaction model. These different aspects lead us to believe that we should take into account different dimensions of trust when designing ACAs.

In this paper, we first propose a new model of trust cognitive mechanisms in the context of interactions between human and artificial conversational agent. We do so by combining multi-disciplinary research on human-to-

Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA
clara.bove@axa.com (C. Bove); clara.bove@lip6.fr (C. Bove); jonathan.aigrain@axa.com (J. Aigrain); marcin.detyniecki@axa.com (M. Detyniecki)
https://kmitd.github.io/ilaria/ (J. Aigrain)

human trust, trust in organizations and trust in technologies. Then, we derive a design heuristic for building trust in ACAs from this model. Overall, we argue that taking into account the diversity of trust drivers is key to have a holistic approach of trust in ACAs.

2. State of the Art

2.1. Dimensions of Trust

The meta-analysis conducted by Lee and See [4] based on 14 researches on trust integrates these parameters and summarizes them in three similar basis needed for building trust in a context of an interaction between an individual and a machine: *Performance*, *Process* and *Purpose*. These 3 concepts can be linked to 3 dimensions of trust for ACAs: technological, oganizational and human aspects of trust.

2.1.1. Technological aspect of trust

Performance represents the competence or expertise of a machine through its actions. Through its performance, a machine demonstrates that it is competent to perform the actions expected and credible for an individual [5, 4, 2]. In addition, performance must be robust and measurable over time [6].

2.1.2. Organizational aspect of trust

Process represents the transparency of a service that must make clear to an individual how it operates. According to Lee and See [4], process is evaluated in terms of consistency, transparency and integrity [2].

2.1.3. Human aspect of trust

Purpose represents the overall intentions of interactions in a given relationship [4]. Purpose is measured by its kindness, honesty and integrity to reduce the feeling of being vulnerable [4, 2].

2.2. Evolution of Trust

Trust is built over time and through interactions, evolving between different stages [7, 8]: trust based on *deterrence*, trust based on *knowledge*, and trust based on *identification*.

2.2.1. Trust based on deterrence

Trust based on *deterrence* is the first stage of trust building and is driven by the concept of punishment. This stage of trust is governed by established and identifiable rules limiting one's vulnerability and so distrust [4]. These rules should not be violated as there is a risk of punishment that can result in trust loss [7].

2.2.2. Trust based on knowledge

The second stage of trust building is based on our knowledge of the trusted object: as an individual's experience with it expands, he/she becomes familiar with it, and his/her knowledge can then be used to assess his/her own vulnerability to the trusted object, predict its behavior and grant it confidence or not [8], adopting an appropriate level of trust [8, 9].

2.2.3. Trust based on identification

The last stage of trust building occurs when the two sides of the interaction understand each other [7]. At this stage of trust, there is no longer a need for a contract or agreement to limit an individual's vulnerability because the person is convinced that the trusted one respects his/her interest and expectations [8].

3. Building Trust

We propose a model of building trust (see Figure 1), based on the dimensions and evolu-

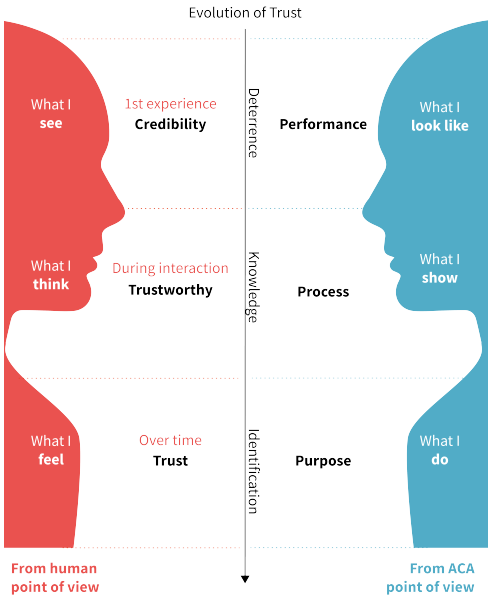

Figure 1: Building trust model for interactions between a human and an Artificial Conversational Agent (ACA)

tion stages of trust. It makes it possible to distinguish human needs and ACA-related characteristics. Each dimension of trust is associated to a stage of Trust evolution [7], Human expectations [2, 4, 5] and ACAs' assets needed to meet them [8, 4, 10]. We also propose a design heuristic (see Table **??**) which translate our model into principles for all NLP practitioners and ACA designers inspired by the heuristic of Amershi et al. [11].

3.1. When first interacting

3.1.1. What a user sees is credibility

The first thing individuals can evaluate is what they **see**. More precisely, an individual makes an initial assessment of what he/she has perceived regarding the ACA, having little or no knowledge of it. Therefore, the evaluation criteria focus first on the **credibility** of the ACA as seen by an individual with respect to his/her expectations of the ACA's skills.

3.1.2. What an ACA looks like relates to its performance

the ACA must **seem** efficient, capable and reliable regarding an individual's expectations for him/her to see it as credible and interact with it. So the first asset an ACA must develop is its **performance** [Principle P1]. The agent's first interactions and actions will allow the user to judge whether it is able to meet his/her expectations. Also, its sentences and actions must demonstrate that it is reliable. Appearances can play an important role as well in the perception of performance and they should not be neglected in any aspect of the first image sent to the user [5, 12, 13].

3.1.3. The interaction is subject to the deterrence effect

When first interacting with an ACA, a user needs **credibility**. The principle of trust based on **deterrence** helps him/her to measure its performance in the absence of knowledge, and protects from vulnerability. If an ACA fails at delivering what is expected by user at first, distrust and neglect are more likely to occur.

3.2. During interactions

3.2.1. What a user thinks is trustworthy

Through his/her interactions with an ACA, a user experiences its actions, anticipate them and react appropriately in the event of errors. As a result, a user **thinks** whether the service is **trustworthy** because the history of their interactions has allowed him/her to understand the process and to know the limits of the conversation.

Context	Principle	
1st experience	P1	**Make ACA credible** ACA's performance appears at first sight in line with to users' expectations, speaks clearly and answers as expected by users.
During interaction	P2	**Make ACA trustworthy** ACA's processes are transparent and users are able to understand how the conversation is handled by the ACA and how decisions or actions are taken.
	P3	**Manage error scenarios** Possible errors (such as misunderstanding, out-of-topic requests, sensitive topics or negative feedbacks) are anticipated and ACA manage each of them in order to minimize users' frustrations.
Over time	P4	**Build a trust-based relationship** ACA is emphatic and anticipate users' needs. It generates affect and identification though conversations.

Table 1
4 design principles for trusted human-ACA interactions, categorized by context of use

3.2.2. What an ACA shows is its process

For a user to perceive trustworthiness in ACA and engage with it, he/she needs transparency about the process it uses to generate content and make decisions, so that he/she can measure the risk of each interaction. Thus, ACAs' services need to **show** transparency over the **process** [Principle P2]. Interpretability solutions could be leveraged when a Machine Learning model is used to provide further transparency [14, 15, 11, 16]. Also, he/she needs to understand the limits of the conversation to avoid situations of misunderstanding and frustration. As for transparency over the process, ACAs' services need to manage error scenarios [Principle P3] to remain trustworthy for the user.

3.2.3. Knowledge for appropriate level of trust

A user can only objectively judge whether the ACA is **trustworthy** if he/she has **knowledge** on it. If the ACA shows a sufficiently transparent process and limits frustrations, then this individual will be able to make an informed decision as to whether an AI is trustworthy, adopt the appropriate level of trust and thus develop his/her relationship with it.

3.3. Over time

3.3.1. What a user feels is trust

Until this stage, the user remained in control of all interactions and can continue to interact without fearing for his/her vulnerability. But once the ACA has accumulated enough data about him/her and begins to make decisions, the user begins to lose control over the interactions and it can increase his/her vulnerability if there is no trust. Therefore, it is important that the individual **feels** safe regarding the ACA service to be willing to **trust** and accept the decisions it makes.

3.3.2. What an ACA does is its purpose

From the moment a user has confidence, he/she is more willing to give up some amount of control. Beyond uses and interactions, he/she agrees to rely on the decisions of an ACA if he/she is convinced of its kindness, honesty and integrity. Thus, the last asset that

ACA must put forward to gain the trust is the **purpose** [Principle P4]. We believe an ACA should provide the required information so that a user can be convinced of its benevolence.

3.3.3. Emotional identification

The last stage of trust development is more subjective because it implies that an individual no longer needs to give consent to let the ACA act. This feeling, which is specific to **identification**, develops if the individual is personally convinced of the other party's good **purpose**.

4. Conclusion and Considerations

Since trust is key to our relationship with ACAs, it is essential to understand its mechanisms so that service designers and practitioners can make good use of it. Our work shows that the dimensions of technological, organizational and human trust presented in previous research are relevant and need to be combined to better understand the cognitive development of trust in the context of ACA. As a technology, ACA needs to have a credible performance in the eye of the user. As a service, ACA's processes need to be clear and transparent in order to be perceived trustworthy and limit frustrations. And as a social agent, ACA needs to have a good purpose to create a safe and trust-based relationship.

Further work will focus on testing this model in real world operations to further evaluate the design heuristic.

References

[1] L. T. Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, R. Atun, Conversational agents in health care: Scoping review and conceptual analysis, Journal of medical Internet research (2020).

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

[3] K. Schaefer, J. Chen, J. Szalma, P. Hancock, A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems, Human Factors: The Journal of the Human Factors and Ergonomics Society 58 (2016).

[4] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80.

[5] B. Fogg, Persuasive Technology. Using computers to change what we think and do, Morgan Kaufmann Publishers,Elsevier, San Francisco, 2002.

[6] T. B. Sheridan, Telerobotics, automation, and human supervisory control, 1993.

[7] R. J. Lewicki, B. B. Bunker, Trust in relationships: A model of trust development and decline, Jossey-Bass, San Francisco, CA, 1995.

[8] C. Corritore, B. Kracher, S. Wiedenbeck, On-line trust: Concepts, evolving themes, a model, International Journal of Human-Computer Studies 58 (2003) 737–758.

[9] K. Hoff, M. Bashir, Trust in automation, Human Factors: The Journal of Human Factors and Ergonomics Society 57 (2015) 407 – 434.

[10] P. Hancock, D. R Billings, K. Schaefer, J. Chen, E. de Visser, R. Parasuraman, A meta-analysis of factors affecting trust in human-robot interaction, Human factors 53 (2011) 517–27.

[11] S. Amershi, D. Weld, M. Vorvoreanu,

A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '19, ACM, New York, NY, USA, 2019, pp. 3:1–3:13.

[12] E. Ostrom, A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, American Political Science Review 92 (1998) 1–22.

[13] R. Pally, Emotional processing: The mind-body connection, The International journal of psycho-analysis 79 (Pt 2) (1998) 349–62.

[14] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[15] C. Baru, Data in the 21st Century, Springer International Publishing, Cham, 2018.

[16] X. Renard, T. Laugel, M.-J. Lesot, C. Marsala, M. Detyniecki, Detecting Potential Local Adversarial Examples for Human-Interpretable Defense, in: Workshop on Recent Advances in Adversarial Learning (Nemesis) of the European Conference on Machine Learning and Principles of Practice of Knowledge Discovery in Databases (ECML-PKDD), 4, Springer International Publishing, Dublin, Ireland, 2018.