Trusting Intelligent Machines

Deepening trust within socio-technical systems

Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart and Simon Wells

Introduction

Intelligent machines have reached capabilities that go beyond a level that a human being can fully comprehend without sufficiently detailed understanding of the underlying mechanisms. The choice of moves in the game Go (generated by Deep Mind's Alpha Go Zero [1]) are an impressive example for an artificial intelligence system calculating results that even a human expert for the game can hardly retrace [2]. But this is, quite literally, a toy example. In reality, intelligent algorithms are encroaching more and more into our everyday lives, be it through algorithms that recommend products for us to buy, or whole systems such as driverless vehicles. We are delegating ever more aspects of our daily routines to machines, and this trend looks set to continue in the future. Indeed, continued economic growth is set to depend on it. The nature of human-computer interaction in the world that the digital transformation is creating will require (mutual) trust between humans and intelligent, or seemingly intelligent, machines. But what does it mean to trust an intelligent machine? How can trust be established between human societies and intelligent machines?

The concept of trust plays an important role in many contexts [3]–[5]. In the social world trust is about the expectation of cooperative, supportive, and non-hostile behavior. In psychological terms, trust is the result of cognitive learning from experiences of trusting behavior with others. Philosophically, trust is the taking of risk on the basis of a moral relationship between individuals. In the context of economics and international relations, trust is based on calculated incentives for alternative behaviors, conceptualized through game theory. Game theory is also used in the field of multi-agent systems to model trust between artificial agents. In the case of management of organizations, trust is about exposing vulnerability, while assuming that the other individual will not take advantage of this. In the world of automation, trust is seen as a feature of entities that can be calculated as the probability of reliable behavior in the presence of externally induced uncertainty. In general, perhaps the common summary of the various views of trust is that it expresses a willingness to take risk under uncertainty [6].

Levels of Trust

There is a long tradition of work looking at trust between humans, or between artificial agents (see, for example, [7]). But to what extent can these results be transferred to

interactions between humans and intelligent machines? Before we can address this question, we need to distinguish between three levels of "trust". We call these *inductive trust*, *social trust*, and *moral trust*.

Inductive trust is derived from personal past experience: a person trusts something because it has previously acted in the way that they expected (e.g. they have used it themselves, they have seen somebody else using it, they have used something very similar, or they have been told by somebody trusted or read from a trusted source what behavior they should expect). This type of trust is the simplest to formalize, being based simply on an estimation of expected outcomes. When the concept of "trust" between humans and machines is discussed, it most commonly inductive trust that is being suggested.

However, in many places the resulting "sense of trust" is misplaced, as humans have a tendency to over interpret the depth of their relationships with machines, and to think of them in social or moral terms where this is not warranted. This is especially likely to occur where the intelligence within the machine is opaque, as this tends to lead people to overestimate the level of intelligence within the machine. In contrast, over-skepticism by humans, i.e. the view that no machine can really be trusted, will prevent the intended use of the technology through unnecessary fear. For example, people tend to prefer human forecasters to algorithmic forecasters, even when they have seen the algorithm outperform a human forecaster [8]. In general, the challenge is to accurately model the risk and consequences of a machine failing to act as anticipated, in order to prevent the interaction between the human and the machine being based on a disproportionate attribution of trust, or indeed, the attribution of too little trust.

The next level of trust between entities, *social trust*, attributes goal-directed behaviour to the machine in the encounter between humans and machines, such that the machine has its own goals that it is trying to achieve in a strategic way, in a similar manner as humans do in the context of their social interactions. Because the goals of two humans, or two machines, or a human and a machine may not be the same, the decision about whether to trust or not involves strategic reasoning by the truster. Given that the entity being trusted is likely to act in a way that furthers its own goals, how should I respond in order to achieve my goals? This can also be formalized using game theory, and has been extensively studied in the multi-agent systems [9], economics [10] and general trust management literature. However, this work has to date focused either on person-person or machine-machine interactions, as opposed to interactions between people and machines.

The third level of trust is *moral trust*, where trust between the interactants is based on a shared sense of rights and obligations, rooted in a principled sense of what is right and wrong, that can override strategic concerns. In human-human interactions, this corresponds to a belief that the person being trusted is benevolent [11]. This trust is based on shared values, i.e. that the person will behave in a good way towards the one that trusts them, beyond the level of goodness implied by the possible benefits that are derived from doing what is expected – e.g. payment. This type of trust has been largely unexplored within AI and computer science, yet is clearly fundamental to human-human relationships. Therefore, we finish by asking how and to what extent this type of trust relationship can be replicated between groups of humans and machines.

In the remainder of this paper we examine how trust can be established and maintained, and how this varies among the three types of trust. We argue that explainable AI plays a key role both in the initial establishment of trust, and in repairing trust relationships that have broken down. A crucial point about the establishment of trust is that a trust relationship between a person and a machine does not exist in isolation. Trust spreads through a network of peers, and is also delegated to higher levels such as firms and institutions. We consider how trust can be formalized and operationalized within an intelligent machine, and the extent to which existing work allows us to do this. We finish by considering the prospects for sharing values and morals between people and intelligent machines, allowing for true moral trust.

Mechanisms for establishing and maintaining trust

Why do we need explainable AI?

Where a system is used routinely, e.g. taking an autonomously driving car to work every day, then explanations are unlikely to be requested by the user on a daily basis. In other words, once inductive trust is established, the daily decision of whether or not to trust the system becomes subconscious. But how can trust between a user and a new intelligent machine become established in the first place? If the machine is new, and has not been used by anyone before, then inductive trust cannot function as there are no past examples to go by. This is a crucial problem for early adopters of a technology -- why should they trust a new machine? Explainable AI holds the key here. If a machine can provide an explanation to the early adopter as to why it is acting in the way that it is, this can give the early adopter a reason to trust it in the absence of past experience. This is because an explanation of how the system will act reduces the risk that the user perceives in the interaction. Note that these types of explanations are *ex ante* -- knowing how something works in advance of using it can help you trust it. This is particularly the case with early adopters who are likely to be tech savvy and hence have some level of domain expertise.

Explainable AI is also needed to repair inductive trust relationships that have broken down, by providing an explanation to the user as to why unexpected behavior occurred. An example of this in our everyday lives occurs when a satellite navigation system sends a driver along a different route than usual, which violates the driver's inductive model of how the system should be behave. Without an explanation for the deviation, the user's inductive trust in the system is likely to be undermined. This is because a small number of occurrences of unexpected behavior is likely to outweigh a much larger number of occurrences of expected behavior when the user is making their trust decision. But a human understandable explanation, e.g. that the normal route is blocked by an accident, allows the user to disregard this counterexample. The role of the explanation is therefore to realign the deviation with the user's underlying model of behavior.

What type of explanation do people need?

There is no generally accepted technical formal definition of the terms *explainability* and *interpretability*, however a common understanding should suffice. An intelligent system is interpretable if it can produce information about why and how it reached a result in a form understandable by people. For example, for a deep learning algorithm used for image recognition, this may mean that the system can explain which part of the input picture contributed most to the result and which did not [12]. If the reasoning process can be made knowable and inspectable, the data rendered understandable, and the path from data to decision made intelligible, then a relationship of trust could be constructed. Given this, understanding a specific decision is akin to understanding the reasoning process, and the specific data that the process has operated upon. In order to aid people to understand decisions, without first becoming experts in specific machine learning techniques or whichever other formalism underpins the machine's decision-making intelligence, an intelligent machine should exploit trust-building exemplars from within the wider world. Explanation is one such exemplar. When people wish to understand the decisions of others then they request an explanation, an account of the reasoning that lead to the decision. Real world explanations are commonly verbal or textual, and occur in a wide range of circumstances. Such circumstances can range from domestic, interpersonal relationships, through public companies explaining poor trading performance, to the various parties to a legal case explaining their behavior with respect to the question at issue.

Furthermore it is likely that trust between people and machines will be improved if the people concerned can understand the decisions that are made [13], and can interpret those decisions within a more familiar framework. Such an approach can contribute not only to the trust relationship but has positive implications for related notions of fairness, unbiasedness, privacy [14] trustworthiness, and understandability [15]. In the case of machine intelligence, an explanation would consist of an account of the process that was followed to get from the incoming data to the decision. This account may incorporate statements about the specific data that did or did not affect the outcome as well as statements that serve to link the data to the outcome, and perhaps statements that serve to limit the scope of the decision. Such an account might incorporate linguistic statements, turning the data into words and phrases, which would be in line with the majority of real-world human-human explanatory practices, but need not be limited to such. Whilst we generally expect a verbal or textual explanation for a given set of circumstances, an explanation can also comprise processes of highlighting, filtering, or otherwise constraining the available data solely to that which is pertinent. An explanation, however rendered, is also an opportunity to introduce ancillary information that perhaps sheds light on the context in which a decision is made. An additional benefit of linguistic explanations is the associated increase in intelligibility and interpretability of the system. Turning data into a concise natural language explanation can turn an otherwise opaque system into one that can be understood and predicted.

An important question when designing the system is who exactly should the explanations be aimed at? Explanations should be given based on the person requesting the explanation. For example, an autonomously driven car may state to its passenger that it has a problem with its engine, while to a mechanic it may state the technical details of the problem. However, the onus is on us as developers to improve education of users about intelligent machines, so that users have the correct conceptual model of what an intelligent machine is in order to avoid them being misled by an explanation, i.e. thinking more is going on in the machine than there actually is. Finally, we note that sometimes domain expertise will be

necessary to be able to derive any meaningful understanding from an explanation. A prime example would be intelligent machines working in the medical domain, where explanations are likely to need to be aimed at medical doctors rather than patients. In such cases the patient trusts the machine because of the trust they have in the doctor.

However, even explanation aimed at domain experts may not always be possible. Intelligent machines are, for specific problems, beginning to exceed human capability. In many cases this is due to the increased ability of machines to work constantly, consistently, at scale, and at speed. In some areas, for example Deep Mind's Alpha Go Zero, the results appear to exceed human ability; some moves are made that are novel and inexplicable to human Go-playing experts and yet are effective, leading to more wins and new insights into the game [16]. This raises the question of the limits of explainability. If a human expert cannot interpret the results, beyond saying that they are correct, then the onus is upon the machine to automatically generate an explanation that is intelligible and sheds light upon the process. This relies upon the assumption that there is nothing supra-human about the reasoning processes involved; i.e. that there are no basic processes occurring within the machine that is beyond human capability. Regardless, how to build an intelligent system that can generate novel, correct, but non-understandable results which are in turn explained and made understandable is a challenging open question. The construction of such a system might be considered to constitute a form of Turing test [17] for explainable machine intelligence.

The problems of explanation and justification become acute when one considers stochastic AI methods. For example, an evolutionary algorithm may search through millions of possible solutions, using stochastic operators such as recombination and mutation to move between solutions. This stochasticity presents issues when trying to recreate decisions at a later date. There exist a number of approaches to making such methods more explainable; for example, saving additional meta-information during the algorithm’s execution may facilitate the construction of an explanation of the final solution. Alternatively, for population based approaches, an explanation that contrasts the chosen solution against other members of the final population may provide some form of justification for the final solution. A more radical approach would be to alter the algorithm itself, only using operators that make decisions which can subsequently be presented to the user in a manner that shows their contribution to the final solution.

Networks and levels of Trust

When we form a trust relation with a single entity, we not only trust actions and information performed by and received from this other entity, but also implicitly establish transitive trust relations with other entities. This serves as a shortcut for the creation of trust relationships. We no longer have to gain direct experience of interacting with an entity directly, in order to slowly build up our own inductive model of its expected behavior. Instead, our existing trust relationships can be exploited (Fig. 1). It is important to stress here that trust is contextual -- I may trust a mechanic recommended to me by a friend, but only if that friend has experience

with cars. In other words, we may trust humans or machines in some contexts but not others, based on our judgment of their competence in that context.

An example of this is when we trust an application because somebody we know (and trust) trusts the application as well. Therefore, we assume the application is trustworthy. The main idea, however, is that we did not generate trust in the application *per se* but trust a person that trusts the application. This effect is likely to be crucial when establishing trust in a new technology: we are likely to use a new technology because our friends also use it. Trust in new technologies therefore propagates from the early adopters that build their own inductive models from personal experience, out through wider social networks by exploiting the transitivity of trust. This serves as a shortcut to save us from having to assess the application ourselves.

Trust networks do not just operate horizontally, in this peer-to-peer manner, but trust relationships can also be displaced vertically from a lower level (1st order trust) e.g. intelligent machines; to a higher level (2nd order trust) entity e.g. companies, or mechanisms such as rules governing interactions [18]. For instance, when an individual buys a car, it does not need to trust each worker that produces each part. Rather, it only needs to trust the firm. Similarly, a user trusts a website because of the sole presence of a small green padlock next to the URL which signals a secure and hence trustworthy site, i.e. the padlock acts as a *trust trigger* [19]. Likewise, a user of an e-commerce website can trust a seller based on its previous reviews because he trusts that the review system itself is not biased or manipulated. The fundamental benefit of such 2nd order trust is that it reduces the number of trust relationships by (i) aggregating numerous entities into one higher entity and by (ii) extending trust from a lifetime-limited entity (humans, machines, etc.) to a virtually unlimited entity (companies, rules, etc.). Overall, this massively reduces the costs of creating, maintaining and monitoring trust relationships.

Networks of trust, whether horizontal or vertical, are ubiquitous. Yet, they also present important challenges. For instance, a minor change in one trust relationship can affect the connected trust relationships and potentially create an avalanche with dramatic effects. If a user realizes that the review system of an e-commerce website is flawed, he might lose trust in any of the sellers on that website. Such challenges have been extensively studied in human-human relationships in economics, and in machine-machine relationships in multi-agent systems. But the case of human-machine relationships in which agents do not necessarily share morals, norms and emotions are still poorly understood. Further work is needed to understand how networks of trust can be operationalized and adapted to human-machine relationships.

Operationalizing Trust

Inductive trust can be operationalized through building statistical models of expected behavior, and explainable AI that establishes initial trust and repairs the trust relationship when it is threatened by unexpected behavior. But how can social and moral trust be operationalized?

Operationalizing social trust through game theory

We need to consider social trust whenever a machine has its own goals that it is acting to fulfill. Because these goals may not necessarily coincide with our own, this involves strategic reasoning to determine whether to trust the machine or not. Game theory provides a formal model for carrying out this reasoning, and has shown that the following mechanisms can support trust relationships:

- **Repeated interactions** (direct experience / personal experience): long-term relationship between the same entities, where actions can be conditioned on past behavior, can lead to reciprocal cooperation as proven by the Folk Theorem of game theory [20]. This means that an AI system needs to be able to take account of past experience.
- **Reputation**: Information about how an artificial system has acted towards other people in the past can be used to decide whether to trust that system or not [21]. AI can be rated by users, and this rating needs to be transparent. But this raises the question of which reputation system/norm should be used, i.e. what counts as trustworthy and untrustworthy actions by the system, and how is reputation updated after each interaction? And when can we trust reputation given that this might be faked, e.g. reviews of a system by fake users?
- **Partner choice:** A marketplace of potential partners encourages those partners to act in a trustworthy way so that other individuals will partner with them in future [22]. In the context of AI, this means that rival AI systems from different producers should be encouraged.
- **Apology/forgiveness**: In the context of human-AI interaction, errors are unavoidable. An AI system can apologize when such errors occur in order to build a long-term relationship with humans [23]. Theory shows that the apology needs to be delivered in a sincere way to be effective, such that it is costly to the producer of the system in material / financial terms [24].
- **Emotions**: Evidence from human-computer interaction studies [25] shows that AI/robots can be trusted more when they are capable of emotional expressions, e.g. feeling guilty when making a mistake. A recent game theory model of guilt shows that internalized norms, such as self-punishment when feeling guilty, can promote cooperation in a population of self-interested agents [26]. In the context of human-AI interaction, AI systems can express their guilt emotion when making an error, for instance by providing a sincere, costly apology or doing self-harm if an apology is not feasible, in order to maintain trust from humans. However, at least for the present, any appearance of emotion necessarily involves deceiving users.

Future work needs to investigate the extent to which results from behavioral economics experiments on person-person interactions can be transferred to machine-person interactions. A key question is how can we prevent mechanisms such as emotions and apologies from being used to deceive users. The development of appropriate ethical rules is needed for this.

Operationalizing moral trust through ethics

Moral trust involves shared values and norms of behavior [27]. This might imply the machine's decision making is based on its own 'moral code'. Alternatively, moral trust could also be established *transitively* with machines, through moral trust in the designer of the machines. As in human-human relationships, the presence of transparent and trustworthy professional ethics is critically important for the trustworthiness of technological systems [28] produced and maintained by these people (e.g. the Hippocratic oath of doctors). This could be operationalized through a set of codified professional ethics for AI that people cannot opt out from using, and inevitably trust [29]. This set of rules must include basic grounding regulations:

- A machine must always identify itself as a machine when initiating a conversation or when challenged to do so, to avoid deceiving users into thinking that the machine is human. There is a risk that this could happen accidently now that we have new realistic speech production technology such as Google Duplex [30].
- Individuals and organizations with legal responsibility for the machine must be identified.
- When providing explanations or justifications of actions a machine must neither deliberately provide false statements nor deliberately withhold relevant information in order to deceive the entity that requested the information.
- Any decision made by an intelligent machine must be repeatable at a later stage. Consequently, program code and data, including context data and metadata delivering information such as origin and ownership, must be archived. When applying stochastic methods pseudo-random number generators must be initialized with known values and these values have also to be archived.

Considering the items in the list above it can be concluded that legal means must be applied, although we note that this itself rests on trust in the institutions of the legal system. Developers must commit to include appropriate announcements in automated conversations and printed statements in written outputs. AI developers must ensure that the mechanisms used are inherently "honest". Note that software systems have been subverted in the past, as in the "Diesel gate" scandal [31]. The challenges of ensuring the proper collection of both code and data admittedly requires considerable technical expertise and effort. However, it may be argued that the scientific community should have overcome such difficulties in the past with the need to ensure that published scientific work is both repeatable and verifiable.

Concluding, programmers and providers of intelligent machines have to provide the principal benevolence of their systems and a degree of transparency that ensures that the working mechanisms of the system can be retraced whenever necessary. We note that the direct transparency of professional ethics may require expert knowledge; however those who lack this can derive the trustworthiness by relying on expert checkers (auditors) and in turn on the easier to understand professional ethics of the auditors.

Artificial social constructivism and the prospect of shared morals between humans and intelligent machines

Can moral trust be established directly with a machine itself, rather than transitively through its designer? We advocate applying sociological [32] and psychological [33] principles to socio-technical systems, for the reason that it allows the computational or digital agents better interpretation of human behavior. This better understanding by digital counterparts in interactions may lead to longer lasting relationships between agents in the system, facilitating the mechanisms for social trust discussed above. But moreover, it opens the possibility for moral trust in the machine itself. Artificial social constructivism is one method by which human values, and by extension morals, can be upheld in a socio-technical system consisting of both humans, and digital or computational agents. Based on Berger and Luckmann's seminal sociological treatise [34], this theory adapts the same principles of social constructivism from a society of purely humans to a mixed society of humans and computers. Humans create their own reality through specific uses and definitions of words and repeated interactions to create a social reality of norms over time, which are then used to inform and define future behavior.

According to artificial social constructivism, values can be established through first establishing and educating agents about system norms. The norms, which are established over time, are passed on to both existing and new agents in the system through education. The method of education can vary, from following a leader, to general observation and learning from previous interactions [35]. The idea is that, similar to the way a child brought up in a certain system of behavior begins to view the system's successes and failures as its own [36], through learning, maintaining and enforcing the norms of the system, the agents become invested in keeping it going and thus more motivated to uphold it as a whole. Thus, the agents establish generalized norms which the whole system will adhere to. The integration of the norms and values in this way upholds the key principles of value-sensitive design [37] and allows the system the potential to develop and maintain the norms, values and morals that are upheld in human society.

Once established, moral trust allows the trust decision to be cut short. Social trust is computationally the most expensive form of trust as it requires evaluating the expected gain from different possible actions that are available to agents in a trust-requiring situation. But if we can assume that the other entities are all following the same norms or moral rules, then this acts as a shortcut that avoids us having to carry out this calculation. This is analogous to the way that inductive trust can be viewed as an instrument that saves the evaluation efforts for a given situation by assuming that it will produce the same outcome as in previous occurrences.

Binmore similarly argues that moral trust in human societies is an efficient shortcut for social trust [38]. He expresses that as "Ethics arose from nature's attempt to solve certain equilibrium-selection problems". Apparently, evolutionary processes have favored certain successful behavioral patterns and implanted these into the gene pool of populations, or are

passed on to following generations by social learning. Figure 2 summarizes the differences between social and moral trust.

Conclusions and Future Directions

In this paper, we have considered the nature of trust with respect to intelligent technology. In particular, we distinguished between three 'levels' of trust: firstly, what we called *inductive trust*, which is based on a mechanistic, perhaps statistical model of expected operation; secondly, *social trust*, which is required to provide predictive leverage and to coordinate expectations among agents that have their own goals; and thirdly, *moral trust*, where the intelligent machine is expected to have a comparable understanding for qualitative human values, such as morality, justice and rights.

However, the significant contribution of this article is to suggest that we have reached what might be conceptualized as an 'ethical crossroads'. There is an essential difference between non-intelligent and intelligent technology, and this is that use of intelligent technology is bi-directional, that is, as the user utilizes the technology, so the technology can influence its user. However, unlike with other media-technology with potential bi-directional properties (e.g. the same concerns were expressed for radio, and television), the actual intentions of the programmer may be less clearly perceived (or more easily hidden) in interactions with intelligent machines than in, say, watching television. It is therefore imperative that the notions of social and particularly moral trust are used for pro-social benefits, and not an illusory front for more insidious motives.

Acknowledgements

This article arose from the Trust in Intelligent Machines workshop held at Edinburgh Napier University, 8th-10th May 2018. This was funded by an award from Edinburgh Napier University's Researcher Development Fund to Simon T. Powers, Cedric Perret, Neil Urquhart, and Simon Wells, who together organized the workshop. The authors would like to acknowledge the academic leadership of Simon T. Powers in this project, and his contribution in taking the lead in writing the article.

[1] D. Silver *et al.*, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.
[2] “How Google’s AI Viewed the Move No Human Could Understand | WIRED.” [Online]. Available: https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/. [Accessed: 14-Aug-2018].
[3] J.-H. Cho, K. Chan, and S. Adali, “A Survey on Trust Modeling,” *ACM Comput Surv*, vol. 48, no. 2, pp. 28:1–28:40, Oct. 2015.
[4] D. Gambetta, “Can We Trust Trust?,” in *Trust: Making and Breaking Cooperative Relations*, D. Gambetta, Ed. Blackwell, 1988, pp. 213–237.
[5] F. Fukuyama, *Trust: The Social Virtues and the Creation of Prosperity*, 1st Free Press Pbk. Ed edition. New York, NY: The Free Press, 1996.
[6] N. Luhmann, *Trust and Power*. Malden, MA: Polity Press, 2017.
[7] S. P. Marsh, “Formalising trust as a computational concept,” University of Stirling, 1994.

[8] B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm aversion: people erroneously avoid algorithms after seeing them err,” *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, Feb. 2015.
[9] M. Wooldridge, *An Introduction to MultiAgent Systems: Second Edition*, 2nd edition. Chichester, U.K: Wiley, 2009.
[10] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.
[11] N. Lankton, D. McKnight, and J. Tripp, “Technology, Humanness, and Trust: Rethinking Trust in Technology,” *J. Assoc. Inf. Syst.*, vol. 16, no. 10, Oct. 2015.
[12] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” *ITU J. ICT Discov. Spec. Issue No 1*, Oct. 2017.
[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, pp. 1135–1144.
[14] D. Kim and I. Benbasat, “The Effects of Trust-assuring Arguments on Consumer Trust in Internet Stores,” in *ICIS 2003 Proceedings*, 2003.
[15] B. Y. Lim, A. K. Dey, and D. Avrahami, “Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2009, pp. 2119–2128.
[16] D. S. Weld and G. Bansal, “The Challenge of Crafting Intelligible Intelligence,” Mar. 2018.
[17] A. M. Turing, “I.—COMPUTING MACHINERY AND INTELLIGENCE,” *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950.
[18] A. Greif, *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge, UK: Cambridge University Press, 2006.
[19] J. Lumsden, “Triggering Trust: To What Extent Does the Question Influence the Answer When Evaluating the Perceived Importance of Trust Triggers?,” in *Proceedings of the 23rd British HCI group annual conference on people and computers : Celebrating people and technology*, Cambridge, UK, 2009, pp. 214–223.
[20] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.
[21] M. Kandori, “Social Norms and Community Enforcement,” *Rev. Econ. Stud.*, vol. 59, no. 1, pp. 63–80, Jan. 1992.
[22] G. Roberts, “Competitive Altruism: From Reciprocity to the Handicap Principle,” *Proc. R. Soc. B Biol. Sci.*, vol. 265, no. 1394, pp. 427–431, 1998.
[23] A. Vasalou, A. Hopfensitz, and J. V. Pitt, “In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions,” *Int. J. Hum.-Comput. Stud.*, vol. 66, pp. 466–480, Jun. 2008.
[24] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts, “Apology and forgiveness evolve to resolve failures in cooperative agreements,” *Sci. Rep.*, vol. 5, Jun. 2015.
[25] C. D. Melo, S. Marsella, and J. Gratch, “People Do Not Feel Guilty About Exploiting Machines,” *ACM Trans Comput-Hum Interact*, vol. 23, no. 2, pp. 8:1–8:17, May 2016.
[26] L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, and T. A. Han, “Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems,” in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, Richland, SC, 2017, pp. 1422–1430.
[27] K. Sood, “The Ultimate Black Box: The Thorny Issue of Programming Moral Standards in Machines [Industry View],” *IEEE Technol. Soc. Mag.*, vol. 37, no. 2, pp. 27–29, Jun. 2018.
[28] “House of Lords - AI in the UK: ready, willing and able? - Artificial Intelligence Committee.” [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm. [Accessed: 14-Aug-2018].
[29] F. Alaieri and A. Vellino, “Ethical Decision Making in Robots: Autonomy, Trust and Responsibility,” in *Social Robotics*, 2016, pp. 159–168.
[30] S. Wong, “Should you let Google’s AI book your haircut?,” *New Sci.*, vol. 238, no. 3178, p. 21, May 2018.
[31] G. Topham *et al.*, “The Volkswagen emissions scandal explained,” *The Guardian*.
[32] J. Pitt, “From Trust and Forgiveness to Social Capital and Justice: Formal Models of Social Processes in Open Distributed Systems,” in *Trustworthy Open Self-Organising Systems*, W. Reif,

G. Anders, H. Seebach, J.-P. Steghöfer, E. André, J. Hähner, C. Müller-Schloer, and T. Ungerer, Eds. Springer International Publishing, 2016, pp. 185–208.
[33] P. R. Lewis, M. Platzner, B. Rinner, J. Tørresen, and X. Yao, Eds., *Self-aware Computing Systems: An Engineering Approach*. Springer International Publishing, 2016.
[34] P. L. Berger and T. Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Penguin UK, 1991.
[35] B. T. R. Savarimuthu and S. Cranefield, "Norm Creation, Spreading and Emergence: A Survey of Simulation Models of Norms in Multi-agent Systems," *Multiagent Grid Syst*, vol. 7, no. 1, pp. 21–54, Jan. 2011.
[36] J. Dewey, *Democracy and Education*. CreateSpace Independent Publishing Platform, 2012.
[37] B. Friedman and P. H. Kahn Jr., "Value Sensitive Design: Theory and Methods," 2002.
[38] K. Binmore, *Natural Justice*. Oxford: Oxford University Press, 2005.

Figure captions

Figure 1: A network of trust is established among trusting entities towards an unknown but trusted entity. The thickness of the arrow indicates the amount of trust. Solid arrows represent direct trust relationships, while dotted arrows represent indirect trust relationships. Here A trusts T directly. Since B trusts A, it also trusts T in a transitive manner. However, the amount of trust is reduced as a result of its indirectness. In a similar way, C, D, and E trust T because B trusts A and A trusts T.

Figure 2: Trust relationships can be positioned along a range bounded by two extreme cases. On one side, moral trust is underlied by internalized mechanisms e.g. biological instincts, internalized norms. The cost to modify or update such mechanism is high. As a result, moral trust is slow to modify but more stable. On the other side, social trust is underlied by externalized mechanisms e.g. rules. Therefore, these mechanisms are easy to modify although at the cost of low stability. The usefulness of moral trust and social trust depends on the context with moral trust being more adapted to stable environments and social trust being more adapted to dynamic environments.

Figures

Figure 1

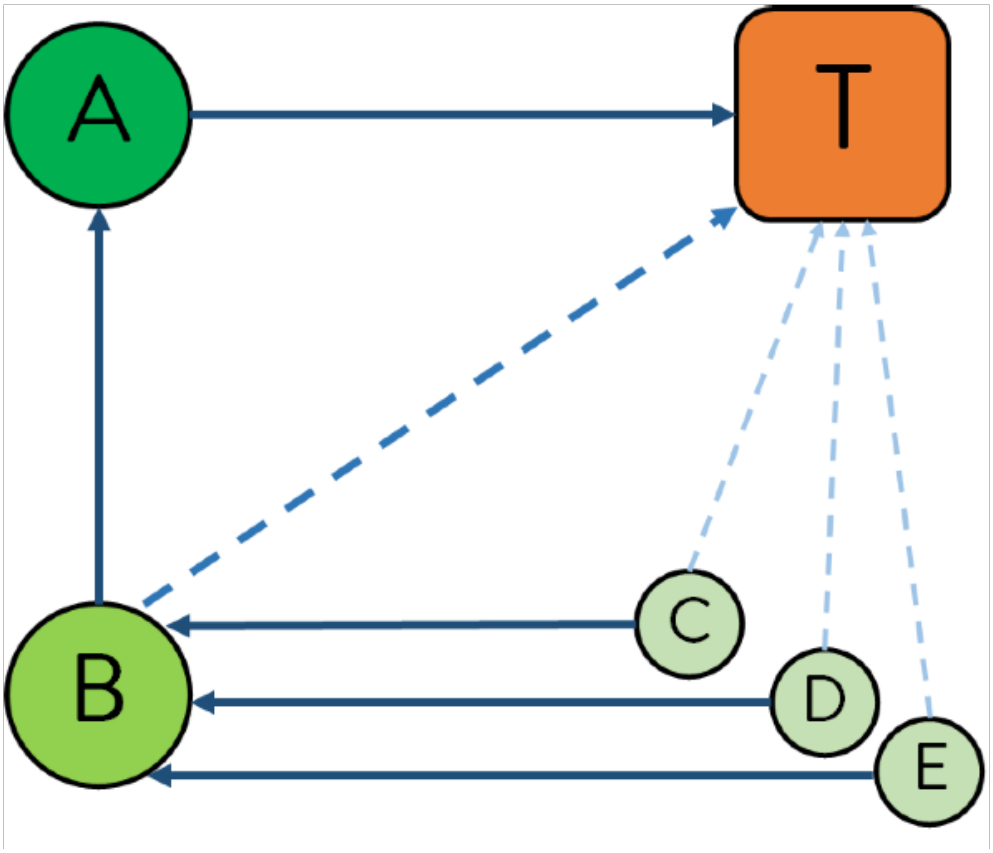
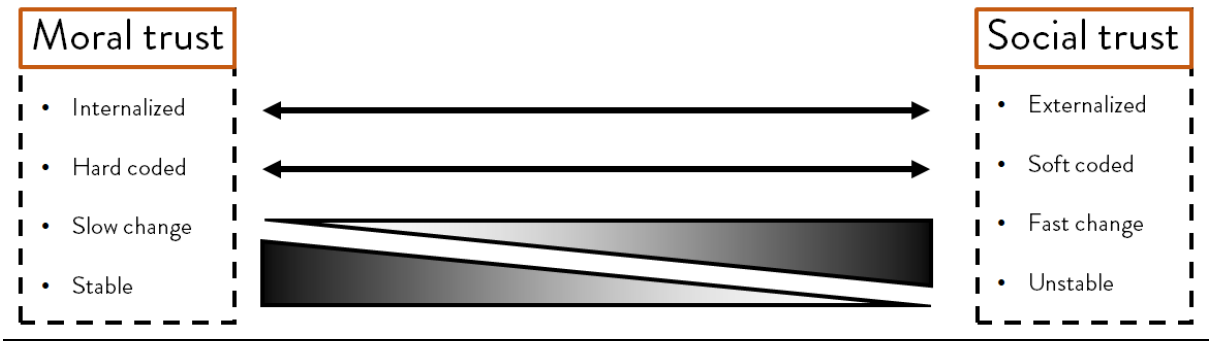


Figure 2


Moral trust
Internalized
Hard coded
Slow change
Stable
Social trust
Externalized
Soft coded
Fast change
Unstable