Trusting in Human-Robot Teams Given Asymmetric Agency and Social Sentience (Extended Abstract)

Geert-Jan M. Kruijff
DFKI GmbH
Saarbrücken, Germany
gj@dfki.de

Abstract

The paper discusses the issue of *trusting*, or the active management of trust [Fitzhugh, Hoffman, and Miller, 2011], in human-robot teams. The paper approaches the issue from the viewpoint of asymmetric agency, and social sentience. The assumption is that humans and robots experience reality differently (asymmetry), and that a robot is endowed with an explicit (deliberative) awareness of its role within the team, and of the social dynamics of the team (social sentience). A formal approach is outlined, to provide the basis for a model of trusting in terms of (i) trust in information and how to act upon that (as judgements about actions and interactions, at the task-level), and (ii) the reflection of trust between actors in a team, in how social dynamics get directed over time (team-level). The focus is thus primarily on the *integration* of trust and its adaptation in the dynamics of collaboration.

Introduction

Trust is dynamic. Whatever exactly trust is *about*, it is something that actors build up, maintain, change over time. It relies on an actively managed process, which is why we use the term *trusting* as proposed in [Fitzhugh, Hoffman, and Miller, 2011].

This paper is about trusting in human-robot teams. We divide (and connect) trusting across task-level, and team-level aspects of teamwork (G. Kaminka).

At the task level, an actor's trust is a weighted expectation about aspects that play a role in determining own behavior. These aspects include expectations about whether another actor will behave in a particular way, whether a particular state will obtain, or whether particular facts (beliefs) can indeed be considered true. What makes trusting complicated here is that we assume experience between humans and robots to be inherently asymmetric. They experience reality differently, which makes it hard to assume an objective notion of truth for information [Kruijff, 2012]. This has an impact on how they can build up common ground [Klein et al., 2004]: What ends up in common ground is not a proposition with an absolute truth value, but an alignment [Pickering and Garrod, 2004] between individual judgments about the "nature" (type) of a particular experience [Cooper, 2012]. Trusting at this level thus becomes trusting judgments, and the alignment of one another's judgments – something which we formally model as (abductive) proofs.

When determining behavior at the task-level, an abductive proof towards an explanation is constructed, of how a specific goal could be reached. These proofs are constructed over situated propositions or meanings. An important component of such situated meaning is the social "context" in which that meaning gets construed [Ginzburg, 2011]. Here, we consider this social context (by itself a situation) to be made up by the social relations between the involved actors, and the physical aspects of where these actors are. Social relations are reflected by roles. A role gives rise to (expectations about) obligations and commitments to specific types of intentions. Together, this intentional aspect of the social context within a situation sets up a facilitating context within which we consider actor abilities and opportunities, i.e. whether or not a social structure facilitates individual actors to act in particular ways.

A failure to perform a particular action may thus give rise to a conflict, if the current social structure does not make it possible for an actor to act such that a joint goal is achieved. The result is a change in social structure, i.e. re-coordination at the team-level. Again we can cast this as an abductive proof, namely as reasoning towards an explanation of how roles and social structure could be restructured in such a way that the task-level goal can be achieved. In essence this is a form of shared control, or adaptive autonomy – but grounded directly in a situated, intentional perspective on interdependence [Johnson et al., 2011] rather than abstract, discrete levels [Parasuraman, Barnes, and Cosenzo, 2007].

Proofs to determine behavior at task-level thus rely on a given social structure; and this social structure acts as an interface with the proofs at team-level, which help to construct / maintain a contextually appropriate (facilitating) structure. Proofs at each of these levels relies on a form of weighted abductive reasoning which can handle uncertainty and incompleteness in information [Kruijff, Janíček, and Lison, 2010; Janíček, 2011]. This is then how trusting comes in: It provides a further form of weighting statements used in proofs. Depending then on the level, we weight expectations, assertions, or facts (task-level), or statements about how to

change social structure given obligations and commitments (team-level). These concepts are discussed in some further detail in the remainder of this paper.

Common Ground, Social Sentience, and Asymmetric Agency

Human-robot teaming is a collaborative, intentional activity. There is a reason why actors act and interact in a particular way. We use abductive inference as the basic mode of reasoning in situated action and interaction [Hobbs et al., 1990; Stone and Thomason, 2003; Kruijff, Janíček, and Lison, 2010]. This type of inference derives an explanation: Δ explains why we believe Γ can happen given our knowledge Σ, i.e. $\Sigma \wedge \Delta \models \Gamma$.

There are two crucial things to observe here. First, we focus explicitly on the *proof* underlying the conclusion, $\Pi[\Sigma \wedge \Delta] \models \Gamma$. The proof steps make explicit *what* information we base the explanation on. Second, the explanation we draw is a *judgment*: We infer that Δ is of a particular *type* t, $\Delta[t]$. As a type it has an internal structure, rather than that it has an objective truth (i.e. a truth value in a model shared by the different actors involved) [Martin-Löf, 1984; Cooper, 2008].

Proofs draw from various sources of information to construct their conclusions. In keeping with the characterization of meaning outlined above, we can see that a proof essentially circumscribes a situation in which a certain set of actions is to be, or has been, performed, to achieve an inferable goal. It appeals to information constituting a focus (relative to which a goal is to be achieved), several resources (beliefs about the world, and what other actors might believe [Lison, Ehrler, and Kruijff, 2010]; existing plans), as well as a dynamic social structure (e.g. knowledge about actions; roles, their needs and obligations [*et al*, 2012]). See also [Kruijff, Janíček, and Lison, 2010; Janíček, 2011] for examples.

As the collaboration progresses, we thus get a sequence of proofs: Proofs explaining how the robot can achieve a particular goal (collaborative action selection and -planning), linked to proofs explaining why a human actor is doing what she is doing (intention recognition). By appealing to situations, these proofs build up a dynamic structure or "universe" over how the robot believes these situations hang together. We can first of all consider this at the level of the dynamics of these situations themselves. Consider σ to be a situation, in the sense of characterizing a focus, a social structure, and (pointers to) reference situations. Furthermore, let α be the non-empty sequence of actions implied by a proof $\Pi[\Sigma \wedge \Delta]$ to help establish the goal $\Delta[t]$. Then, if we understand $\sigma[(\Pi[\Sigma \wedge \Delta])\alpha]$ in the dynamic sense, that is apply the sequence of actions α resulting from Π to (or "in") the situation σ, we should get to a new situation σ' in which the goal $\Delta[t]$ "holds."

More precisely still, the result of the application of α to σ typically is a sequence of situations, of which σ' is only the end-result. And the proof makes explicit, what the information the inclusion of these actions in the inference is based on. Now, given that robots invariably need to act under uncertainty and incomplete knowledge, we need to address this in our inferences. [Kruijff, Janíček, and Lison, 2010; Janíček, 2011] show how uncertainty can be included by constructing a probabilistic version of weighted abduction [Hobbs et al., 1990]. They also show how a basic form of incomplete knowledge can be dealt with through the notion of *assertion*, similar to [Brenner and Nebel, 2009]. An assertion is a (logical, probability-weighted) statement about a piece of information which *is* needed to construct the proof, but for which the robot has neither positive nor negative indications. An example is the assertion that there is a door, to gain access into a building, if the goal is to explore the inside of a building. If this assertion turns out to be falsified (i.e. there is no door), we need to reconsider the course of actions to be taken. In continual planning, assertions are therefore used as explicit points in a plan at which re-planning may be needed.

Here, we suggest to extend the notion of assertion, and the (existentially closed) logical language for constructing proofs with the notion of strong negation [Wansing, 2001]. Whereas the classical notion of negation basically entails a failure to prove, strong negation states something explicitly as not possible or justified. Strong negation has been considered in several approaches to knowledge representation, to include an explicit notion of (closed) falsifiability – which we can now put "opposite to" the notion of assertion as an explicit notion of (open) verifiability. Strong negation says something cannot be the case on the basis of what is known (or the proof fails), where an assertion states that something is assumed to be the case on the basis of what is essentially *not* known (or, again, the proof fails).

If we now look back at our proofs, as judgements anchored to a complex structure over situations, we thus see that with the inclusion of assertions and strong negation we obtain a framework in which we can represent and reason with the asymmetry inherent to a human-robot team. First of all, attributed and shared beliefs become judgements based in proofs which can be qualified with statements about explicit verifiability and falsifiability. That changes these beliefs from "true statements" into subjective judgements about others, presumed to hold under the continual observations of the other's actions. And if a proof turns out to become invalidated (assertion- or strong negation-wise), this is then immediately traceable to the beliefs these proofs are based on, indicating what needs to be retracted.

We can take this a step further though. There is no reason why we can only reason about beliefs, and how these beliefs lead to actions, already observed or observable. We can *lift* verifiability/falsifiability to the level of intentional reasoning, and reason about what we expect to do or not to do, in the light of what is necessary to do.

With the constructions at hand, we can define an additional level of proofs. This level essentially captures the team work. Each proof is cast as a temporal sequence of actions, with accompanying references to situations, and with explicit verifiable/falsifiable references to the achievability of specific goals by (or through) specific agents. These latter goals in and by themselves can again be translated into proofs, anchoring them in the actual situations. This is cru-

cial: It enables to anchor the team work in the ongoing task work set in a dynamic environment, and it makes it possible to reason about how the team can actually achieve its goals together. This leads to a possibility to deal with what we define here as *social sentience*:

> **Social sentience**: *The notion of* social sentience *implies a capability for an actor to reason explicitly with its role within a social structure, how the assumption of this role requires the assumption of certain responsibilities (goals to be achieved) with respect to other roles – and how the inability to fulfill some or all of these responsibilities may require shifting such responsibilities to other actors, resulting in a shift of roles within the social structure.*

Trusting

It is in the context of the above (formal) view that we want to place a basic notion of trusting. Our primary interest is in how we can formulate trusting as a situated-dependent weighting of statements, used in proofs for determining behavior at task- and team-level. This intentional perspective is similar to the trust decision cycle discussed in [Fitzhugh, Hoffman, and Miller, 2011]. For the moment we do not make further distinctions into (externally defined) different types of trust(ing), and instead consider directly their use within proofs.

A proof, as said, is an inference towards an explanation. This typically takes the form of a goal to be achieved – as an explanation for why someone else acted in a particular way (plan/intention recognition), or how to act oneself to achieve this goal (planning). These proofs are based on statements derived from Σ, which themselves are either beliefs about (reported) observations, expectations about commitments and actions given role-based obligations, and assertions about future states. Formally, we can type such statements based on structure and content. Each statement ς gets a weight w to reflect the degree of certainty in this statement. A statement can be based on an expectation or an assertion, to reflect forms of incompleteness in information.

It is straightforward to extend this representationally with a notion of trust as weighting. The weight represents a trust in the source: For example, whether the actor trusts the information provided by another actor, or that another actor will perform an expected action. We add a trust-weight w_t to the uncertainty weighting w_u by constructing a vector $[w_t, w_u]$. We define several functions over this vector: The usual projections π_1, π_2 to provide the individual weights, and a function $f([w_t, w_u])$ over the entire vector to yield a composite weight w. With f we can continue to use weighted abduction as defined in [Janíček, 2011]. At the same time, the projections and the vector make it possible to consider trust separately from uncertainty. Accordingly, proofs can be ranked within a proof space in terms of the composite weighted, and by the individual summations over their separate projections.

This kind of representation is not novel. Its combination with the different kinds of proofs we construct, at task- and at team-level, does provide several novel lines of investigation though. Each derive from the question, how we get at the w_t values.

One interesting aspect here is to consider w_t to reflect *character type*, as a combination of the agent type logic defining how e.g. obligations, commitments, or information from other agents are handled (cf. [Cohen and Levesque, 1990; van Linder, van der Hoek, and Meyer, 1998]), and a multi-dimensional (discrete scale) characterization of interpersonal relationship values [Sheldon, 2004]. We can connect sub-logics from different agent types to intervals on the character trait scales, using e.g. lattice structures over sublogics to ensure consistent composite models; and, connect the traits to different types of trust, for example the ones as presented by [Fitzhugh, Hoffman, and Miller, 2011]. In this way, trust-as-a-weighting arises from a more complex set of character traits, which have a direct influence on how the actor actually *decides* to behave towards the other actor. Both the process of proving, and the resulting proofs themselves, are affected.

Another aspect concerns the grounding of how a particular trust-as-weight comes about, *in particular situations*. As discussed above, proofs deal with judgments, and these judgments are based in a (complex) notion of situated meaning – including typical spatiotemporal situations, as well as more abstract resources (background knowledge), and social context. This gives rise to the possibility to base trust on judgment. Proof universes develop over time, as a reflection of gaining more, and more precise information. Following a continual paradigm in proving-for-planning, acting in real world scenarios is thus likely to revise proofs into new instances. This can reflect both a growth and a development of connected judgments, all within against the background of asymmetric agency and social sentience. With improving or diminishing success in achieving goals, traces through an unfolding proof universe can influence trust between actors, grounded within the particular (social) situation in which the trace is placed. For example, a robot may decide to put less faith in a Mission Specialist's judgments about what he reports to see when he is operating under a significant cognitive workload, in a smokey environment – but more when everything is quiet and easy-to-observe.

Conclusions

The paper considers possibilities how to ground a notion of trusting, as a form of actively managing "trust" between actors in a team, in a framework for situated common ground that starts from asymmetric agency, and social sentience. From the viewpoint of modeling trust, a known idea is used: represent trust as a weighting, to direct what kinds of proofs or plans an actor makes. It is then the connection to the kinds of proofs an actor makes, that opens up new possibilities for modeling trusting: Grounding "it" (or rather, character traits giving rise to trust-as-a-weighting) in proof traces, and then having "it" reflected through selection of different sublogics to guide how information and action with respect to another are reasoned with, within proofs to decide how to act (or interpret another's actions). We thus suggest to see trust and trusting less as standalone concepts, and more to

deal with them directly in terms of how they can guide actor behavior.

Acknowledgments

The work reported in this paper was supported by the EU FP7 Integrated Project NIFTi, "Natural human-robot coordination in dynamic environments" (Grant #247870). For more on NIFTi, see `http://www.nifti.eu`. The author would like to thank Jeff Bradshaw for discussions.

References

Brenner, M., and Nebel, B. 2009. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems* 19(3):297–331.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Cooper, R. 2008. Records and record types in semantic theory. *Journal of Logic and Computation* 15(2):99–112.

Cooper, R. 2012. Type theory and semantics in flux. In Kempson, R.; Asher, N.; and Fernando, T., eds., *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier.

et al, G. K. 2012. Experience in system design for human-robot teaming in urban search & rescue. In *Proceedings of Field and Service Robotics (FSR) 2012*.

Fitzhugh, E.; Hoffman, R.; and Miller, J. 2011. Active trust management. In Stanton, N., ed., *Trust In Military Teams*, Human Factors in Defence. Ashgate. 197–218.

Ginzburg, J. 2011. Situation semantics and the ontology of natural language. In Portner, P.; Maierborn, C.; and von Heusinger, K., eds., *The Handbook of Semantics*. de Gruyter. 830–851.

Hobbs, J.; Stickel, M.; Appelt, D.; and Martin, P. 1990. Interpretation as abduction. Technical Report 499, AI Center, SRI International, Menlo Park, CA, USA.

Janíček, M. 2011. Abductive reasoning for continual dialogue understanding. In *Proceedings of the ESSLLI Student Session 2011*.

Johnson, M.; Bradshaw, J.; Feltovich, P.; Hoffman, R.; Jonker, C.; van Riemsdijk, B.; and Sierhuis, M. 2011. Beyond cooperative robotics: The central role of interdependence in coactive design. *IEEE Intelligent Systems* 81–88.

Klein, G.; Feltovich, P.; Bradshaw, J.; and Woods, D. 2004. Common ground and coordination in joint activity. In Rouse, W., and Boff, K., eds., *Organizational Simulation*. New York City, NY: John Wiley. 139–184.

Kruijff, G.; Janíček, M.; and Lison, P. 2010. Continual processing of situated dialogue in human-robot collaborative activities. In *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN 2010)*. IEEE.

Kruijff, G. 2012. Achieving common ground under asymmetric agency and social sentience in communication for human-robot teaming. In *Proceedings of the 10th IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR 2012)*.

Lison, P.; Ehrler, C.; and Kruijff, G. 2010. Belief modelling for situation awareness in human-robot interaction. In *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN 2010)*.

Martin-Löf, P. 1984. *Intuitionistic Type Theory*. Napels, Italy: Bibliopolis.

Parasuraman, R.; Barnes, M.; and Cosenzo, K. 2007. Adaptive automation for human-robot teaming in future command and control systems. *International Journal of Command and Control* 1(2):43–68.

Pickering, M., and Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27:169–225.

Sheldon, L. 2004. *Character Development and Storytelling for Games*. Course Technology. Cengage Learning.

Stone, M., and Thomason, R. 2003. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*.

van Linder, B.; van der Hoek, W.; and Meyer, J.-J. 1998. Formalizing abilities and opportunities of agents. *Fundamenta Informaticae* 34(1,2):53–101.

Wansing, H. 2001. Negation. In Goble, L., ed., *The Blackwell Guide to Philosophical Logic*. Cambridge, MA: Basil Blackwell Publishers. 415–436.