Using Doctrines for Human-Robot Collaboration to Guide Ethical Behavior*

Geert-Jan Kruijff & Miroslav Janíček
German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{gj, miroslav.janicek}@dfki.de

Introduction

Considerations of risks and ethical dimensions of robot behavior have so far been largely restricted to the issue of governing lethal behavior in military robotics, see e.g. Arkin (2007), Lin et al. (2008). Within this field, the importance of such considerations is hardly underestimated:

> "It is essential that, before unmanned systems become ubiquitous (if it is not already too late) that we consider this issue [of the ethics of robot lethal behavior] and ensure that, by removing some of the horror, or at least keeping it at a distance, we do not risk losing our controlling humanity [...]."
> DCDC/MOD (2011)

These issues are real. Robots with lethal capabilities are being deployed in several operational theaters. And mistakes are being made. Partly they are due to "fog of war" and the difficulty for remote operators to have a good operational awareness of the situation on the ground. But, as anecdotal evidence in for example Singer (2009) points out, another reason is unclarity in command, control and communication (C3) between operators, robots, and those in the field – an unclarity that gets compounded as all these actors are often geographically dispersed. And then, people die.

We would like to argue that these issues concern us all. Most researchers eschew (direct) military research, and that might be a reason for why the issue of ensuring ethical behavior in robots has been largely marginalized in the fields of robotics, and human-robot interaction. Put somewhat over-simplified, "We don't make robots that kill, we don't want to, hence this doesn't concern us." But this just isn't correct.

Even when we disregard the impact of military robotics on the public perception of robots, risks and ethical considerations of right and wrong robot behavior go well beyond the military domain. We can find ample of examples in domains we (the authors) ourselves are active in: Urban Search & Rescue, and robots for assisting children undergoing medical therapy. What if an autonomous rescue robot ignores the detection of a victim – and the victim turns out later to have needed immediate assistance? What if the robot reacts inappropriately to a child's emotional state – and the child becomes uncooperative towards further treatment? Sharkey (2008) discusses similar issues, pointing out the current ethical frontiers in human-robot interaction. The point is, any real-life application in which robots act together with humans, there are real risks and "rights and wrongs" to be considered. Seriously.

How serious this can get is illustrated by the results of a study into acceptability of lethal robots, reported in Moshkina and Arkin (2007). These results show a tendency for people to find more autonomous robots *capable of lethal behavior* less acceptable. More acceptable are robots that are extensions of a soldier, fully under his/her control (i.e. tele-operated). In the tele-operated case, the results show the operator (soldier) to be the most responsible for any lethal errors made by a robot under his control. Yet, as soon as a robot gains a higher degree of autonomy, the results show a shift. Forty percent of the respondents in this study would assign blame to the robot – and more importantly, people see this shift to concern both the robot and its designer.[1] Now consider this observation in the light of the fact that, when it comes to robots for e.g. Urban Search & Rescue, the drive is towards *more* autonomy, not less – see e.g. Birk and Carpin (2006), Murphy et al. (2008b). Dealing appropriately with risks and ethical behavior in *non*-military settings might thus actually prove to be quite important. But how?

Various authors have looked at the issue of ethical behavior from a design perspective, for example Wallach and Allen (2009) and most notably Arkin (2007, 2009). They look specifically at how the behavior of an individual robot can be guided (or even controlled) ethically, given a basis of ethical behavior as specified by e.g. the international Laws of War and mission-specific Rules of Engagement.

*The research reported here was financed by the EU FP7 ICT Programme, Cognitive Systems Unit, project "NIFTi: Natural Human-Robot Cooperation in Dynamic Environments" (#247870). URL: [http://www.nifti.eu]

[1] "People" here covers military personnel, policy makers, scientists, and general public – on the average with a high education.

We do not dispute the importance of imbuing robots with a moral sense, to drive their autonomous behavior such that it stays within the realms of the ethically right. However. Robots do not act fully autonomously, nor do they act in isolation. They act in the context of a team, together with humans, and other robots. The vision is for humans and robots work together as an organic asset. Hence we would like to argue that as such it is not only the behavior of any team individual that is (to be) judged ethically. It ultimately comes down to the behavior of the entire team – and that is, in turn, based in how cohesive a unit it manages to be.

Unit cohesion plays an important role in individual performance ("well-being"), and resilience in dealing with performing under stress; Siebold (2000), MHAT-7 (2011). Conversely, lack of unit cohesion (and, closely tied to that, leadership) often provided for a context in which unethical behavior could arise, or at least where "wrong behavior" could arise; See Singer (2009) and Arkin (2010). A question is thus how, in a human-robot team, we should design robots and their behavior to optimally contribute to unit cohesion.

A proper model of command, control, and communication (C3) provides the basis for successfully building up units as teams. This is what is generally understood as the *doctrine* of how a unit is to operate, cf. e.g. the U.S. Army's *Mission Command: Command and Control of Army Forces* (FM 6-0, 2003), or for first responders, the FEMA *Field Operations Guide for National Urban Search and Rescue Response System* (US&R-2-FG, 2003). As Siebold (2000) discusses, interaction patterns between members of a team play an important role in this. They interact with group development, and their effect on performance is that they contribute to social bonding and norms. And this effect ties our focus on doctrines back to the ethical issue.

Our focus here is on interaction patterns, within the context of building up and maintaining unit cohesion. Appropriate interaction patterns for a human-robot team aim to facilitate optimal behavioral controls by providing clear ways of command, control, and communication mediated by a common ground (situation awareness) among the team members. At the same time, appropriate communication can aid to meliorate negative effects of circumstances on behavior (notably, stress). In this setting, we consider a *communication doctrine* to be a specification of how team members are to communicate among themselves: With whom, about what, when, and in what role. An outline of our approach is as follows.

Outline of the Argument

We consider human-robot teams as a socio-technological complex. Humans, robots, multi-modal interfaces are all tightly interwoven into a complex system. It is this system as a whole that acts, that brings about situation awareness, that achieves goals in reality. It is this system that ultimately needs to behave according to a set of rules that "we" consider to be "good." Our goal in this paper is to see how we could build up a formal model of human-robot teaming, which would make it possible *for a robot*

1. to use that model to track the social dynamics of the team;
2. within that context, to contribute to a team behavior that is "good," (and not, through action or inaction, move towards "bad" behavior); and,
3. to make its actions fully accountable.

Naturally, this is a tall order. We will not solve the problem in this paper. But we do hope to provide an argumentation, a basis from which we can view the issue of human-robot teaming in the light of ethical behavior. This first of all requires us to clarify the philosophical views we adopt.

When talking about "ethical behavior" we intuitively mean "good behavior." What exactly needs to be understood by good behavior we leave aside for the moment. Typically this is a matter of the kind of setting in which a human-robot team is to be deployed, and the cultural backgrounds of the team itself, and the area where it is deployed. We leave that aside. Instead, we consider what it means to say that some agent behaves well. The "well" here is a judgment. It is a judgment we feel we can make because the agent can be held *responsible* for his or her behavior.

This raises several questions, first of all *what* it means to hold someone (morally) responsible, and *why* you can do so. Watson (1996) discusses two alternative views on moral responsibility: Responsibility as *attribution*, and responsibility as *accountability*. Attribution involves a dependence on judgment of character, or characteristic attributes of the agent: "he *behaved* like that, for he *is* like that;" See e.g. Smith (2008). Accountability includes an explicit element of volition, a deliberate choice made by the agent: "he *behaved* like that, because he *chose* to do so." The responsibility of the agent for his actions is then grounded in the choices he made. The choices make him "blameworthy," as Levy (2005) argues, irrespective of whether we would call the agent good or bad. In this paper, we adopt a view that leans more towards accountability. Within the rules laid out for behaving within a team context, we want the robot to be able to be accountable for its actions. In that sense, we consider the robot responsible, and behaving ethically.[2]

That we even *can* hold a robot responsible is a direct consequence of the socio-technological view on human-robot teams adopt. This view is based on Actor Network Theory; See e.g. Latour (1987), Callon and Law (1997), Latour (2005) (ANT). ANT provides a systemic perspective on the interconnections between heterogeneous technical and non-technical elements. These in-

[2]The resulting issue of normativity then is not so much concerned with assigning blame, but rather, with ensuring compliance.

terconnections yield a network between these elements. This network shapes how the elements act, and interact, and how elements mutually influence each other. Crucial to ANT is the assumption of a radical symmetry in agency between the different kinds of elements in a network: Everything is an actor, whether human or robot. Given interconnectivity, and the resulting dynamics of mutual influence, radical symmetry yields a *relational* notion of agency.

This has important consequences for *how* to consider responsibility. As a relational property of a network of humans and robots, agency and the effects it brings about through behavior imply a notion of responsibility that goes beyond the individual actor. Responsibility rests in part with the individual actors, but ultimately is defined in terms of accountability for the interaction between the actors that resulted in the actions of the actors. Even if we would argue that responsibility in decision-making stays with humans, the mutual influence between human and robot in shaping agency cannot be taken out of the equation. Parasuraman et al. (2007), Parasuraman and Hancock (2008) observe for example how over-reliance on adaptive automation (like robots) influences the behavior of humans, effecting the overall outcome of the team's efforts. Through the dynamics of the social network of a human-robot team, humans shape the agency of robots, and robots shape the agency of humans. Responsibility is about how these dynamics between humans and robots develop, how they can be guided, and how they can be made accountable for (notably, by the robots themselves).

We base these dynamics in a notion of *trust*. If a human trusts in the capabilities of the robot to achieve a particular result, he or she can let the robot do so. And vice versa, if the robot trusts in the information or in the instructions it gets from the human, it can act on it. Without trust in capabilities and intentions, we assume there to be no dynamics. On the other hand, if there are capabilities in which trust can be placed, be they human or robot capabilities, delegation and cooperation of tasks based in these capabilities becomes possible. This is how we can model shared control or adjustable autonomy in a human-robot team; See e.g. Falcone and Castelfranchi (2000, 2001b). Guiding behavior within a team then comes down to controlling in what capabilities trust can be developed, and the effects of such placements of trust on levels of autonomy (i.e. the degrees of agency of the robot).

Given the systemic perspective on the social dynamics of a human-robot team, we phrase these capabilities from the viewpoint of *roles*. A role defines a particular set of functions an actor can provide within a team. Within the context of that role, each function has a particular autonomy "bandwidth" (levels of autonomy associated with the execution of that function, in the sense of Parasuraman et al. (2000)), and a set of integrity limits which constrain the conditions under which the function can be executed. By a capability we understand such a function, which an actor performs in its capacity of playing the role within which this function (in its particular form) is available; See also Burke et al. (2004), Murphy and Burke (2010).

Finally, we can then link this back to the issue of guiding behavior by defining the roles humans and robots can play within a team, i.e. what is normally understood by a "doctrine." Here, we consider the use of design patterns for social interaction as proposed by Kahn et al. (2008).

Background

Several taxonomies of human-robot teams have been proposed to classify the types of interaction that can take place, see. e.g. Yanco and Drury (2004). These taxonomies describe a range of factors including temporal and spatial aspects of the interaction, what roles a human might play, team composition, and the balance between attainable levels of robot autonomy and the amount of human intervention. With these, Yanco & Drury identify several important dimensions that potentially influence the dynamics of human-robot teaming. At the same time, although the dynamics of adaptive autonomy are presented as the defining factor for interaction, these taxonomies yield little insight in what this dimension might actually mean for the roles a robot can actively play in a human-robot team.

We therefore adopt an alternative approach, using a role-based model of human-robot teams. Humans and robots can play different roles, and associated with each role are available levels of autonomy, information, authority, perspectives. This model is based on work by Burke et al. (2004) and Murphy and Burke (2010). Murphy & Burke consider a role as a complex of tasks (what needs to be done), strategies (how things are to be done), and capabilities (knowledge and skills required for the tasks). A team member can play multiple roles, concurrently or sequentially, and always adopts a specific view (on the world, and on information) that is inherent to that role. This viewpoint defines the scope and form of the information that that generates within that role. Murphy & Burke provide several examples of instantiations of human-robot teams, analyzing communicative patterns between team members in terms of roles Burke et al. (2004), Murphy et al. (2008a). Murphy & Burke point out that improving autonomy in perception and behavior might help reduce the number of humans needed in a human-robot team; cf. also Birk and Carpin (2006). However, their discussions lack a (formal) account of how we could properly model the dynamics of adaptive autonomy within roles.

Parasuraman et al. (2000, 2007) present a model that we use to make the notion of (robot) autonomy more explicit. The purpose of the model is to give a description of the degrees to which a system could automate particular processes. Parasuraman et al consider four basic functions, or types of processes, to which automation can be applied in a system: 1) information acquisition, 2) information analysis, 3) decision and action

selection, and 4) action implementation. A system may implement different degrees of automation / autonomy across these process types. We suggest to use a modified form of Parasuraman et al's model to make possible levels of autonomy within a specific role clear. The modifications concern the need to distinguish between perceptual, and informing & decision-making autonomy as suggested in Murphy and Burke (2010). Parasuraman et al. (2000) only model decision- and action selection.

We combine Murphy & Burke's communication-oriented notion of role, with Parasuraman et al's notion of adaptive autonomy, and a notion of joint activity Klein et al. (2004). This helps us to see a *team communicative process* as the combination of communicative functions *about information* and roles that *provide and act upon information*; cf. also Burke et al. (2004). A communicative process clarifies the possible information flows within a team, and how integrity limits and dynamics of adaptive autonomy can influence this flow.

At the same time, this does not yet make clear how communication takes shape. It defines the strategies, but not yet the operational characteristics. We adopt a notion of communicative interaction design pattern based on Kahn et al. (2008) to do so. A communicative interaction design pattern makes explicit the exact integrity constraints (preconditions or triggers), autonomy, acts, and commitments it presupposes for a communicative act between two team members. A pattern thus goes beyond a dialogue act in that it considers the situated and social dimension of communication, beyond the basic function it plays in furthering a dialogue.

Distributed Situation Awareness

A human-robot team is a complex socio-technical system. Its goal is typically to build up and maintain a form of situation awareness (SA) about the environment it operates. This SA serves the coordination and collaboration within the team, as well as the building up of knowledge for an overall assessment of the situation, to serve (future) operations outside the scope of the team. The question is, what shape this SA takes.

Salmon et al. (2009) provide an extensive overview of approaches to modeling SA. Most approaches consider SA to be a notion that applies to an individual actor. It is a cognitive construct inside the head of an individual; See e.g. Endsley (1995, 2000). The problem with such a view is that it reveals little about how such SA would be built up and maintained.

With the ever increasing use of teams in complex systems, new approaches to modeling *team* SA have been proposed, e.g. Salas et al. (1995). Team SA is typically characterized as a common picture arising from the sum of the individuals' SA, the explicitly shared SA, or the overlap between individuals' SA. Aside from the difficulty of assessing such a notion of SA, Salmon et al note (p.32), that models of team SA still lack proper descriptions of the process of how team members build up team SA, and what factors affect team SA. Specifically, interaction between team members appears to be largely ignored.

Furthermore, seeing team SA as a common picture adhered to and relevant for all, appears to be difficult to apply to settings where the actors of a team are geographically dispersed. And this is typically the case for a human-robot team. For example, in an Urban Search & Rescue mission, the robots are typically deployed inside the hotzone, human operators requiring line-of-sight might be inside or near the hotzone, and everyone else is located remotely. There is no strict need to share all the information. Some information remains "local" between the operators and robots, just focusing on the information needs they have to do what they are tasked to do. This may even be restricted to individuals: A robot may but need not convey every bit of sensor information it has to the user. This typically varies with the (granted) level of autonomy. This scoping of information similarly applies across operators and robots. Coordination does require information to be shared. But it does not require all information to be at that level.

The alternative to team SA-as-a-common-picture is a more *distributed* notion of situation awareness (DSA). Salmon et al consider how DSA arises from the interactions and relationships between the different aspects of a socio-technical system. DSA "emerges" so to speak from what happens between the elements. Stanton et al. (2006) define it as "activated knowledge for a specific task, at a specific time within a system" (p.1291). What information is required for SA depends on the goals, actions, and actors involved, at a particular point in time. Sharing SA, or rather intentions, information, arises out of the need of individual actors within the system while interacting with others. This drives the updating of (D)SA: actors with limited or degraded SA interact with others to update it, to be able to achieve their goals; See Stanton et al. (2006). Propositional networks are proposed to describe this form of situation awareness; See e.g. Salmon et al. (2008, 2009). A propositional network is a network with directed, named links between propositions. The individual propositions represent concepts, whereas the names on the links between propositions represent relationship-types. It depends on the activity, or a set of activities specific to a particular event or phase during an operation, which parts of such a network get activated. Stanton et al. (2008) explicitly link this activation of parts of a proposition network to a social network for a team, communication patterns, and coordination demands. This indicates how information was provided, by whom, who used it, and why.

The resulting view on situation awareness is thus a systems construct, not an individual cognitive construct. Situation awareness is what happens between the actors, to make it possible to find a compatible level of understanding the environment on which they can act. It relies on an integration of social structure, in-

formation, and communication- and coordination patterns. It is this background against which we develop the role-based perspective, in the next sections.

Social Dynamics in Teams

How do we ensure that, whatever roles a robot plays, it acts as part of a team – irrespective of its level of autonomy? Klein et al. (2004) describe ten challenges for making automation a "team player" in joint activity. We summarize these challenges, and see what requirements they raise for collaborative role models for robots in C2 activities. Klein et al define joint activity as "an extended set of actions that are carried out by an ensemble of [agents] who are coordinating with each other." Klein et al. (2004), Clark (1996). This involves four basic requirements, being: (1) a Basic Compact, (2) mutual predictability, (3) mutual directability, and (4) common ground. We discuss these requirements below, and the specific challenges that Klein et al note.

The *Basic Compact* is an often tacit agreement between the participants in the team to facilitate coordination, work towards reaching a shared goal, and prevent or repair breakdowns in team coordination. A basic compact is not fixed, as Klein et al note: "[p]art of achieving coordination is investing in those actions that enhance the Compact's integrity as well as being sensitive to and counteracting those factors that could degrade it." Klein et al. (2004)(p.91) Klein et al formulate two challenges directly related to the Basic Compact: An agent must understand the Basic Compact, so that it can deliberately act on it (Chl. 1), and it must be able to adequately model other agents' intentions and actions relative to the Basic Compact (Chl.2). This raises several requirements for collaborative role models for robots.

- **Scope of Action**: A role needs to define what actions it performs in information- and decision-management processes, (relative to one or more C2 activities). This outlines the scope of the contributions this role can make to the overall effort, as per the Basic Compact.
- **Bandwith of Autonomy**: A role needs to define the lower- and upper limits on its autonomy in acting in information- and decision-management processes (LOA, Parasuraman et al. (2000)).
- **Integrity Limits**: Complementary to the LOA bandwidth for specific actions a role needs to define a notion of "integrity limits." These need to describe the limits to which contributions can be made, and provide for contingency management e.g. through strategies for direct reporting to identified (active) roles in the team ("who should know when I fail").

In highly interdependent settings like a team, the behavior of the individual actors needs to be transparent to others (*mutual predictability*). It needs to be clear why someone is performing a particular action, so that outcomes and follow-up behavior can be predicted. This is necessary for interdependent action and coordination to be efficiently plannable and executable. Challenges here include predictability itself (Chl.3), the ability of an agent to make pertinent aspects of his actions and intentions clear to others (Chl.5), and the ability to observe and interpret such signals from others (Chl.6).

- **Direct Reports on Acting: To**: A role needs to define strategies for reporting its actions and its reasons for performing them, to one or more roles in a team. This is connected to the Bandwidth of Autonomy, and the issue of keeping the human in the loop.
- **Direct Reports on Acting: From** A role needs to define conditions that reflect interdependence of its own role on others, in terms of information- and decision-making state.

Mutual directability involves both the capability to assess and modify the actions of other agents, and to be responsive to the influence of others on one's own actions. This involves the ability to be directable (Chl.4), to negotiate goals (Chl.7), and to support a continual approach to collaboration (Chl.8). The latter reflects the need to allow for the Basic Compact to be collaboratively adjusted, as the situation demands. Finally, actors must be able to participate in managing attention (Chl.9).

Again, from the viewpoint of collaborative role models and keeping the human in the loop, this requires roles to combine an upper limit to autonomy in decision processing, with reporting (and negotiating) on decision status, as indicated above. Furthermore, a role needs to make dynamic and role-based authority explicit.

- **Authority** A role needs to make explicit authority relations: from what roles it accepts directability, what goals it can negotiate with other roles (and when), and for what actions (states) it has authority to decide (which may include actions associated with other roles). Authority may be inherent to the role itself, or be derived from the adoption of a role in the current team context (e.g. being designated team leader over other agents).
- **Attention** A role needs to define strategies for directing the attention of others in specific roles, on the basis of signals, activities and changes in its own state. Any attention-direction action needs to be mediated through the viewpoints associated with the roles in question.

Finally, effective coordination between team members requires them to establish and maintain a form of common ground, DSA capturing (activated) knowledge, beliefs, and intentions.

- **Viewpoints** A role needs to define its own viewpoint(s) Murphy and Burke (2010) and the modalities through which these viewpoints are mediated.

The above requirements indicate how we could make the inherent interdependence between roles more explicit, in terms of acting assumptions about authority & level of decision-making autonomy, viewpoint & situation awareness, information management between roles, and coordination (including correction, confirmation, negotiation, and attention-directing). This is an abstract, complex characterization to aid us in formulating collaborative role models for robots, interacting in teams with other robots and humans.

We formulate a role as a tuple (*PossibleActors, Functions,Relations,Viewpoints*). *PossibleActors* is a set of actors (types) that can take on this role. *Functions* is a set of function primitives. The function primitives identify what they operate on (information INFO, actions ACT) , and (where relevant) with respect to what (ENVIRONMENT), or whom (the role's primary actor SELF, or other actors ACTOR). ACTORs are identified by roles. ACTs have assigned ACTORs including SELF. Table 1 lists the primitives we consider. The list is an extension of the RASAR-CCS scheme presented in Burke (2003). For a function, the role also specifies a bandwidth of autonomy indicating what level(s) of autonomy are required for performing this function, and which integrity limits condition the possible execution of the function. *Relations* is a set of directed links between the role, and other roles. Each link indicates how an actor "playing" this role can construct connections to other roles active in the team. Finally, *viewpoints* specify from what (mediated or non-mediated) perspectives an actor in this role perceives the situation; See Murphy and Burke (2010).

	Function	**Explanation**
1.	SEEKINFO	Ask for INFO from an ACTOR
2.	REPORT	Share INFO about SELF, ENVIRONMENT, or other ACTOR
3.	CLARIFY	Make previous INFO more precise
4.	CONFIRM	Affirm previous INFO, or (selected) ACT
5.	CONVEYUNC	Express doubt, disorientation, or loss of confidence in INFO
6.	PROVIDEINFO	Share INFO other than REPORT, either in response to a SEEKINFO request from another ACTOR, or to provide unsolicited information
7.	PLAN	Project future, spatially situated GOALS, or ACTs to GOALS
8.	SELECT	Select ACT
9.	EXECUTE	Execute ACT
10.	ORDER	Authority: Order another ACTOR to ACT, or allow another ACTOR to order SELF
11.	INTERVENE	Authority: Allow another ACTOR to intervene in ACT
12.	PROPOSE	Propose ACT(s) to ACTOR

Table 1: Function primitives for robot collaborative role models; (1–7 from Burke et al/RASAR-CCS).

Modeling Dynamics of Trust in Teams

We adopt the logic of trust developed by Herzig et al. (2010). Herzig et al present a formalization of Falcone and Castelfranchi (2001a)'s notion of occurrent trust: The trust of one agent in another agent, for the latter to achieve an actively intended goal, here and now. The formalization is based in a logic of time, action, beliefs, and choices. Beyond occurrent trust, Herzig et al also show how a dispositional notion of trust can be formalized: Trusting that an agent will achieve an intended goal in a (finite) future from now.

The logic defines several operators. $\texttt{After}_{i:\alpha}\phi$ means that "immediately after agent i does α, it is the case that ϕ" (and hence, $\texttt{After}_{i:\alpha}\bot$ means i cannot do α). $\texttt{Does}_{i:\alpha}\phi$ means "agent i is going to do α and ϕ will be true afterwards," with $\texttt{Does}_{i:\alpha}\top$ meaning i will do α. $\texttt{Bel}_i\phi$ represents that i believes ϕ, and $\texttt{Choice}_i\phi$ means "agent i has the chosen goal that ϕ." Using these operators, together with temporal operators $\texttt{G}\phi$ ("ϕ is always true") and $\texttt{F}\phi$ ("ϕ will be true"), Herzig et al then define capability and intentionality as follows: $\texttt{Capable}_i(\alpha) = \neg\ \texttt{After}_{i:\alpha}\bot$, and $\texttt{Intends}_i(\alpha) = \texttt{Choice}_i\ \texttt{Does}_{i:\alpha}\top$.

Together with a Kripke-style model theory and an axiomatization this yields a basic logic in which we can talk about trust. The axiomatization includes a notion of weak realism, stating that $\texttt{Bel}_i\phi \rightarrow \neg\texttt{Choice}_i\neg\phi$ and a principle of intentional action: $\texttt{Does}_{j:\alpha}\top \leftrightarrow \texttt{Capable}_j\alpha \wedge \texttt{Intends}_j\alpha$. This results then in the following definition of *occurrent trust* :

$$\texttt{OccTrust}(i,j,\alpha,\phi) = \texttt{Choice}_i\texttt{F}\phi \wedge \texttt{Bel}_i(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\phi)$$

Thus, i trusts j to do α with respect to ϕ if and only if i wants ϕ to be true at some point in the future and believes j will ensure ϕ by doing action α. Given the definition of the `Capable` operator, the above definition implies that i believes that j is capable of achieving ϕ. Given our model of roles as discussed above, we can apply the notion of occurrent trust as follows. We model functions as actions α, and LOA and integrity limits as conditions on capability (particularly, functional restrictions on the accessibility function for the model-theoretic specification of this operator), getting $\texttt{Capable}_i(\alpha, LOA, limits)$. We adapt the principle of intentional action accordingly, to have `Does` reflect LOA and the integrity limits as conditions.

Herzig et al also define a notion of dispositional trust. This notion is defined with respect to a weaker concept of goal, namely a *potential goal*. A potential goal ϕ in a context κ is a goal which the agent may entertain at some point in the future: $\texttt{PotGoal}_i(\phi,\kappa) = \texttt{Poss}_i\texttt{F}^*(\kappa \wedge \texttt{Choice}_i\texttt{F}\phi)$. Then,

$$\texttt{DispTrust}(i,j,\alpha,\phi,\kappa) = \texttt{PotGoal}_i(\phi,\kappa) \wedge \texttt{Bel}_i\texttt{G}^*((\kappa \wedge \texttt{Choice}_i\texttt{F}\phi) \rightarrow (\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\phi))$$

Similarly, this definition can be enhanced to explicitly reflect autonomy and integrity, as per above.

Herzig et al prove that the resulting notions of trust are related: $\texttt{DispTrust}(i,j,\alpha,\phi,\kappa) \wedge \texttt{Choice}_i\texttt{F}\phi \wedge \texttt{Bel}_i\kappa \rightarrow \texttt{OccTrust}(i,j,\alpha,\phi)$, i.e. given a disposition of i to trust j to achieve ϕ under conditions κ, once i believes κ obtain and i makes an explicit choice for ϕ, i trusts j do achieve ϕ now.

Naturally, this is a simplification: for j to act upon i's choice for ϕ requires communication. The operators do not make this explicit. To do so requires bringing them in line with the DSA viewpoint: intentions, beliefs, choices need to be communicated between the actors involved, given the roles they play. Once they are shared (in the sense of DSA), they facilitate the move from dispositional to occurrent trust. This remains an issue to be addressed, though.

For the moment, we use Herzig et al's logic as a vehicle to see how trust between actors in a social system, mediated by their roles, can be based on beliefs, intentions, choices, and capabilities – with an understanding of capabilities as functions. A notion of *accountability* then comes within reach: Accountability is the requirement of a proof, against this logic, of how trust and interactions between actors led to a specific actor to perform an action, achieving a result ϕ. Given that trust is based in capabilities, and how these capabilities can enter into choice and intention, leads to a way in which accountability (as a form of compliance) can be controlled. Namely, by restricting the availability and nature of capabilities afforded within a human-robot team, particularly on the side of the robot. This is nothing new, per se: If you do not want a robot to shoot innocent bystanders, do not make it possible for it to shoot in the first place.[3] But, if we place them within the notion of roles, DSA, and a social network-based perspective on human-robot teams, we (arguably) get the notion of team-based responsibility we are looking for. A doctrine is a set of capabilities, available within roles, together defining the ways actors within the team can (better: should) interact to maintain cohesion, and behave in a way that ethical.

Conclusions

This paper argued for looking at ethical behavior in human-robot teams from a systemic viewpoint. A team is a system, its behaviors arise at the interactions between the actors within that team. And as such, issues of guiding behavior, accountability, and responsibility take on such a relational nature as well. We discussed how we could make behaviors explicit in terms of roles within a team, associating with each role sets of capabilities and conditions under which these capabilities could be used to perform specific tasks. We then moved to a theory of trust to see how choice, intention, and beliefs between agents could result in agents to trust each other to act, and achieve specific results. Combined with the role-based view, this gives a basis to consider accountability within teams: To hold an actor accountable means there is the possibility to trace back the trust through social interactions that altogether led to an observed state of affairs.

At the same time, the model we discussed here is a simplification. Actions do not always yield the intended results (stochasticity). And neither robots nor humans always have a perfect, encompassing view of the situation in which they act (uncertainty, incompleteness). These factors (stochasticity, uncertainty, incompleteness) need to be factored into a theory of trust, through the logical formalization of beliefs, intentions, and choices, to yield a more realistic account. (And it is at this point that we can get back to views on moral responsibility: none of the mentioned factors are *choices*, they are *judgments*. Hence, it can be expected that ultimately, we need a notion of moral responsibility for human-robot teams that integrates both views or aspects; See also Smith (2008).)

[3] Or, if that somehow fails, ground them: `http://www.wired.com/dangerroom/2008/04/armed-robots-st/`

References

R.C. Arkin. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Technical Report GIT-GVU-07-11, Georgia Tech, 2007. Prepared for the US Army Research Office.

R.C. Arkin. *Governing Lethal Behavior in Autonomous Systems*. Taylor & Francis, 2009.

R.C. Arkin. The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, 9(4):332–341, 2010.

A. Birk and S. Carpin. Rescue robotics - a crucial milestone on the road to autonomous systems. *Advanced Robotics*, 20(5):595–605, 2006.

J.L. Burke. Moonlight in Miami: A field study of human-robot interaction in the context of an urban search and rescue disaster training response training exercise. Master's thesis, Department of Psychology, College of Arts and Sciences, University of South Florida, September 2003.

J.L. Burke, R.R. Murphy, M. Coovert, and D. Riddle. Moonlight in Miami: An ethnographic study of human-robot interaction in USAR. *Human Computer Interaction*, 19((1–2)):85–116, 2004.

M. Callon and J. Law. After the individual in society: Lessons on collectivity from science, technology, and society. *Canadian Journal of Scoiology*, 22(2):165–182, 1997.

H. Clark. *Using Language*. Cambridge University Press, 1996.

DCDC/MOD. The UK approach to unmanned aircraft systems. Technical Report . Joint Doctrine Publication 2/11, Ministery of Defence, Development, Concepts and Doctrine Centre (DCDC), 2011.

M.R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.

M.R. Endsley. Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley and D. J. Garland, editors, *Situation awareness analysis and measurement*. Lawrence Erlbaum, 2000.

R. Falcone and C. Castelfranchi. Grounding autonomy adjustment on delegation and trust theory. *Journal of Experimental and Theoretical Artificial Intelligence*, 12:149–151, 2000.

R. Falcone and C. Castelfranchi. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer Academic Publishers, 2001a.

R. Falcone and C. Castelfranchi. The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems, Man, and Cybernatics, Part A: Sysems and Humans*, 31(5):406–418, 2001b.

A. Herzig, E. Lorini, J.F. Hbner, and L. Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18 (1):214–244, 2010.

P. H. Kahn, N. G. Freier, T. Kanda, H. Ishiguro, J. H. Ruckert, R. L. Severson, and S. K. Kane. Design patterns for sociality in human robot interaction. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278, 2008.

G. Klein, D.D. Woods, J.M. Bradshaw, R. Hoffman, and P. Feltovich. Ten challenges for making automation a team player in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, November-December 2004.

B. Latour. *Science in Action: How to Follow Scientists and Engineers Through Society*. Open University Press, Milton Keynes, 1987.

B. Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, 2005.

N. Levy. The good, the bad, and the blameworthy. *Journal of Ethics and Social Philosophy*, 1(2):1–16, 2005.

P. Lin, G. Bekey, and K. Abney. Autonomous military robotics: Risk, ethics, and design. Technical report, CalPoly, 2008. Prepared for the US Department of Navy, Office of Naval Research.

MHAT-7. Mental Health Advisory Team (MHAT) 7, Operation Enduring Freedom 2010. Report, Office of the Surgeon General, U.S. Army Medical Command, February 2011.

L. Moshkina and R.C. Arkin. Lethality and autonomous systems: Survey design and results. Technical Report GIT-GVU-07-16, Georgia Tech, 2007. Prepared for the US Army Research Office.

R.R. Murphy and J.L. Burke. The safe human-robot ratio. In M.J. Barnes and F. Jentsch, editors, *Human-Robot Interactions in Future Military Operations*, Human Factors in Defence, pages 31–49. Ashgate, 2010.

R.R. Murphy, K. Pratt, and J.L. Burke. Crew roles and operational protocols for rotary-wing micro-UAVs in close urban environments. In *Proceedings of the ACM/IEEE Conference on Human-Robot Interaction (HRI'08)*, Amsterdam, The Netherlands, 2008a.

R.R. Murphy, S. Tadokoro, D. Nardi, A. Jacoff, P. Fiorini, H. Choset, and A.M. Erkmen. Search and rescue robotics. In B. Siciliano and O. Khatib, editors, *Springer Handbook of Robotics*, pages Part F, 1151–1173. Springer Verlag, 2008b.

R. Parasuraman and P.A. Hancock. Mitigating the adverse effects of workload, stress, and fatigue with adaptive automation. In P.A. Hancock and J.L. Szalma, editors, *Performance Under Stress*, Human Factors in Defence, pages 45–58. Ashgate, 2008.

R Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans*, 30:286–297, 2000.

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

E. Salas, C. Prince, P.D. Baker, and L. Sherstha. Situation awareness in teams. *Human Factors*, 37:123–136, 1995.

P.M. Salmon, N. A. Stanton, G. H. Walker, D. Jenkins, C. Baber, and R. McMaster. Representing situation awareness in collaborative systems: A case study in the energy distribution domain. *Ergonomics*, 51:367–384, 2008.

P.M. Salmon, N.A. Stanton, G.H. Walker, and D.P. Jenkins. *Distributed Situation Awareness: Theory, Measurement, and Application to Teamwork*. Human Factors in Defence. Ashgate, 2009.

N.E. Sharkey. The ethical frontiers of robotics. *Science*, 322: 1800–1801, 2008.

G.L. Siebold. Military group cohesion. In T.W. Britt, C.A. Castro, and A.B. Adler, editors, *Military Life: The psychology of serving in peace and combat*, pages 185–201. 2000.

P.W. Singer. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. Penguin Books, 2009.

A.M. Smith. Control, responsibility, and moral assessment. *Philosophical Studies*, 138:367–392, 2008.

N. A. Stanton, R. Stewart, D. Harris, R. J. Houghton, C. Baber, R. McMaster, P. M. Salmon, G. Hoyle, G. H. Walker, M. S. Young, R. Dymott, and D. Green. Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. *Ergonomics*, 49(12-13):1288–1311, 2006.

N.A. Stanton, C. Baber, and D. Harris. *Modelling Command and Control: Event Analysis of Systemic Teamwork*. Human Factors in Defence. Ashgate, 2008.

W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

G. Watson. The two faces of responsibility. *Philosophical Topics*, 24(2):227–248, 1996.

H.A. Yanco and J.L. Drury. Classifying human-robot interaction: An updated taxonomy. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, 2004.