16. Making Agents Acceptable to People

Jeffrey M. Bradshaw[1], Patrick Beautement[2], Maggie R. Breedy[1], Larry Bunch[1], Sergey V. Drakunov[3], Paul J. Feltovich[1], Robert R. Hoffman[1], Renia Jeffers[1], Matthew Johnson[1], Shriniwas Kulkarni[1], James Lott[1], Anil K. Raj[1], Niranjan Suri[1], and Andrzej Uszok[1]

[1] Institute for Human and Machine Cognition
University of West Florida, USA
[2] QinetiQ, Malvern Technology Centre, UK
[3] Tulane University, USA

Abstract

Because ever more powerful intelligent agents will interact with people in increasingly sophisticated and important ways, greater attention must be given to the technical and social aspects of how to make agents acceptable to people [16.72]. From a technical perspective, we want to help ensure the protection of agent states, the viability of agent communities, and the reliability of the resources on which they depend. To accomplish this, we must guarantee, insofar as is possible, that the autonomy of agents can always be bounded by an explicit enforceable policy that can be continually adjusted to maximize the agents' effectiveness and safety for both human beings and computational environments. From a social perspective, we want agents to be designed to fit well with how people actually work together. Explicit policies governing human-agent interaction, based on careful observation of work practice and an understanding of current research in the social sciences and cognitive engineering, can help assure that effective and natural coordination, appropriate levels and modalities of feedback, and adequate predictability and responsiveness to human control are maintained. These factors are key to providing the reassurance and trust that are the prerequisites to the widespread acceptance of agent technology for non-trivial applications.

16.1 Introduction

Since the beginning of recorded history, people have been fascinated with the idea of non-human agencies.[1] Popular notions about androids, humanoids, robots, cyborgs, and science fiction creatures permeate our culture, forming the backdrop against which software agents are perceived. The word robot, derived from the Czech word for drudgery, entered public discourse following Karel Capek's 1921 play *RUR: Rossum Universal Robots* [16.21] (Fig. 16.1).

[1] Works by authors such as Schelde [16.80] and Clute and Nicholls [16.26], who have chronicled the development of popular notions about androids, humanoids, robots, and science fiction creatures, are a useful starting point for agent designers wanting to plumb the cultural context of their creations. Lubar's chapter "Information beyond computers" in [16.64] provides a useful grand tour of the subject. See Ford, Glymour, and Hayes [16.40] for a delightful collection of essays on android epistemology.

While Capek's robots were factory workers, the public has also at times embraced the romantic dream of robots as "digital butlers" who, like the mechanical maid in the animated feature *The Jetsons* would someday putter about the living room performing mundane household tasks (Fig. 16.2).[2] Despite such innocuous beginnings, the dominant public image of artificially intelligent creatures has often been more a nightmare than a dream. Would the awesome power of robots reverse the master-slave relationship with human beings (Fig. 16.3)?[3] Would seeing the world through the eyes of agents lead to dangerously distortions of reality (Fig. 16.4)? Everyday experiences of computer users with the mysteries of ordinary software, riddled with annoying bugs, incomprehensible features, and dangerous viruses reinforce the fear that the software powering autonomous creatures would pose even more problems. The more intelligent the robot, the more capable of pursuing its own self-interest rather than that of its human masters (Fig. 16.5); the more human-like the robot, the more likely it is to exhibit human frailties and eccentricities (Fig. 16.6). Such latent images cannot be ignored in the design of software agents-indeed, there is more than a grain of truth in each of them!

[2] It is interesting to note that today's robotic vacuum cleaners have little resemblance to mechanical maids. However, that is true in part because they are conceived as inexpensive single-function appliances and not multi-purpose assistants. Were our current technical prowess sufficient to build cheap, smart, and versatile robotic assistants, there is little doubt that we would prefer models that featured a "good brain and an unspecialized body" [16.68]. See also Sect. 16.4.2 below.

[3] Whether or not such futures are plausible is besides the point—there is no doubt that the fears are real for many people right now. For example, Bill Joy notes the "prophecy" of the Unabomber, asserting that while his "mentality was criminal, his vision is rather realistic": 'What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.' *Theodore Kaczynski* — the criminal Unabomber. On the other hand, just one year ago Stephen Hawking, the noted physicist, suggested using genetic engineering and biomechanical interfaces to computers in order to make possible a direct connection between brain and computers 'so that artificial brains contribute to human intelligence rather than opposing it.' The professor concedes it would be a long process, but important to ensure biological systems remain superior to electronic ones. "In contrast with our intellect, computers double their performance every 18 months," he told Focus magazine. "So the danger is real that they could develop intelligence and take over the world." [16.55].

Fig. 16.1. Scene from Capek's play *Rossum Universal Robots*

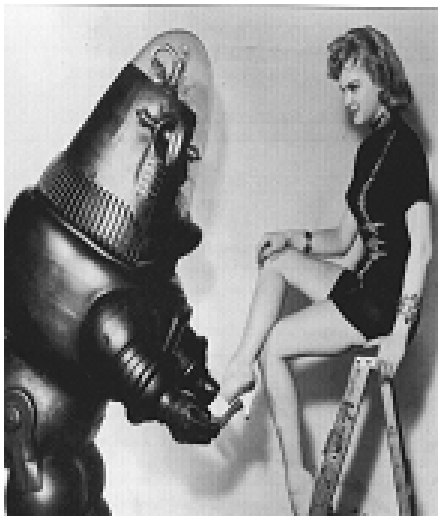

Fig. 16.2. Electro the Robot (aka Robby the Robot) as digital butler to Anne Frances

Fig. 16.3. Powerless in the grasp of a robot (From *Astounding Science Fiction,* October 1953)

Fig. 16.4. Select-O-Vision

Fig. 16.5. A robot thief

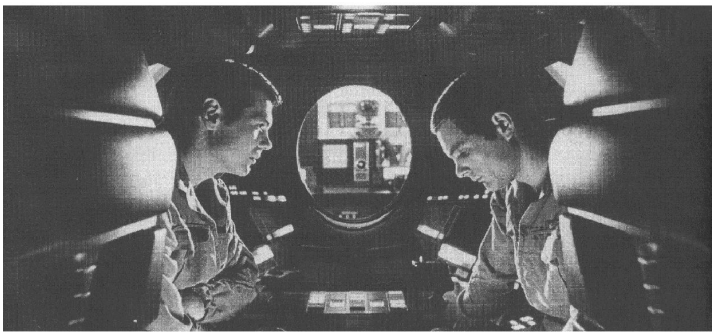

Fig. 16.6. In Arthur C. Clarke's *2001: A Space Odyssey*, human beings are compelled to hide from the psychotic computer HAL

"Agents occupy a strange place in the realm of technology," summarizes Don Norman, "leading to much fear, fiction, and extravagant claims" [16.72]. By their ability to operate independently without constant human supervision, they can perform tasks that would be impractical or impossible using traditional software applications. On the other hand, this additional autonomy, if unchecked, also has the potential of effecting severe damage if agents are poorly designed, buggy, or malicious. Because ever more powerful intelligent agents will increasingly differ from software that people are accustomed to, we need to take into account social issues no less than the technical ones if the agents we design and build are to be acceptable to people. Continues Norman:

The technical aspect is to devise a computational structure that guarantees that from the technical standpoint, all is under control. This is not an easy task.

"The social part of acceptability is to provide reassurance that all is working according to plan.... This is also a non-trivial task" [16.72].[4]

This chapter summarizes our efforts to address, through a policy-based approach (Sect. 16.2), some of the technical and social aspects of agent design for increased human acceptability. From a technical perspective, we want to help ensure the protection of agent states, the viability of agent communities, and the reliability of the resources on which they depend. To accomplish this, we must guarantee, insofar as is possible, that the autonomy of agents can always be bounded by explicit enforceable policy that can be continually adjusted to maximize the agents' effectiveness and safety for both human beings and computational environments (Sect. 16.3).

From a social perspective, we want agents to be designed to fit well with how people actually work together. Explicit policies governing human-agent interaction, based on careful observation of work practice and an understanding of current research in the social sciences and cognitive engineering, can help assure that effective and natural coordination, appropriate levels and modalities of feedback, and adequate predictability and responsiveness to human control are maintained (Sect. 16.4). In short, interaction among humans and agents must be graceful and should enhance rather than hinder human work. All these factors are key to providing the reassurance and trust that are the prerequisites to the widespread acceptance of agent technology for non-trivial applications.[5]

[4] Similarly, Alan Kay has written: "It will not be an agent's manipulative skills, or even its learning abilities, that will get it accepted, but instead its safety and ability to explain itself in critical situations. At the most basic level the thing we want most to know about an agent is not how powerful it can be, but how trustable it is" [16.56].

[5] A more complete study of many of these topics can be found in [16.11, 16.112]. For an entertaining and informative general characterization of various approaches to human-centered computing, see [16.51].

16.2 Addressing Agent Acceptability Through the Use of Policy

The idea of building strong social laws into intelligent systems can be traced at least as far back as the 1940s to the science fiction writings of Isaac Asimov [16.6]. In his well-known stories of the succeeding decades, he formulated a set of basic laws that were built deeply into the positronic-brain circuitry of each robot so that it was physically prevented from transgressing them. Though the laws were simple and few, the stories attempted to demonstrate just how difficult they were to apply in various real-world situations. In most situations, although the robots usually behaved "logically," they often failed to do the "right" thing, typically because the particular context of application required subtle adjustments of judgments on the part of the robot (e.g., determining which law took priority in a given situation, or what constituted helpful or harmful behavior).[6]

Shoham and Tennenholtz [16.84] introduced the theme of social laws into the agent research community, where investigations have continued under two main headings: *norms* and *policies.* Drawing on precedents in legal theory, social psychology, social philosophy, sociology, and decision theory [16.119], *norm-based* approaches have grown in popularity [16.9, 16.33, 16.62, 16.63]. In the multi-agent system research community, Conte and Castelfranchi [16.32] found that norms were variously described as constraints on behavior, ends or goals, or obligations. For the most part, implementations of norms in multi-agent systems share three basic features:

1. they are designed offline, or
2. they are learned, adopted, and refined through the purposeful deliberation of each agent; and
3. they are enforced by means of incentives and sanctions.

Interest in *policy-based* approaches to multi-agent and distributed systems has also grown considerably in recent years (http://www.policy-workshop.org). While sharing much in common with norm-based approaches, policy-based perspectives differ in subtle ways. Whereas in everyday English the term *norm* denotes a practice, procedure, or custom regarded as typical or widespread,

[6] In an insightful essay, Roger Clarke explores some of the implications of Asimov's stories about the laws of robotics for information technologists [16.25]. Weld and Etzioni [16.120] were the first to discuss the implications of Asimov's first law of robotics for agent researchers. Like most norm-based approaches described below (and unlike most policy-based approaches), the safety conditions are taken into account as part of the agents' own learning and planning processes rather than as part of the infrastructure. In an important response to Weld and Etzioni's "call to arms," Pynadath and Tambe [16.76] develop a hybrid approach that marries the agents' probabilistic reasoning about adjustable autonomy with hard safety constraints to generate "policies" governing the actions of agents. The approach assumes a set of homogeneous agents who are motivated to cooperate and follow optimally-generated policies.

a *policy* is defined by the American Heritage Online dictionary as a "course of action, guiding principle, or procedure considered expedient, prudent, or advantageous." Thus, in contrast to the relatively descriptive basis and self-chosen adoption (or rejection) of norms, policies tend to be seen as prescriptive and externally-imposed entities. Whereas norms in everyday life emerge gradually from group conventions and recurrent patterns of interaction, policies are consciously designed and put into and out of force at arbitrary times by virtue of an explicitly-recognized authority.[7] These differences are generally reflected in the way most policy-based approaches differ from norm-based ones with respect to the three features mentioned above. Policy-based approaches

1. support dynamic runtime policy changes, and not merely static configurations determined in advance;
2. work involuntarily with respect to the agents, that is, without requiring the agents to consent or even be aware of the policies being enforced, thus aiming to guarantee that even the simplest of agents comply with policy; and
3. wherever possible, are enforced preemptively, preventing in advance buggy, poorly designed, unsophisticated, or malicious agents from doing harm, rather than rewarding them or imposing sanctions on them after the fact.

In the following subsections, we define policy in the sense that it is used in this section 16.2.1 and distinguish it from related concepts 16.2.2. We then offer definitions of the two major types of policy 16.2.3, describe the relationship between autonomy and policy 16.2.4, discuss both traditional focus areas and new challenges for policy management 16.2.5, and outline the most important aspects and benefits 16.2.6.

16.2.1 What Is Policy?

In agent and distributed computing contexts, policy can be defined as *an enforceable, well-specified constraint on the performance of a machine-executable action by a subject in a given situation.*

- *enforceable:* In principle, an action controlled by policy must be of the sort that it can be prevented, monitored, or enabled by the system infrastructure;
- *well-specified:* Policies are well-defined declarative descriptions;
- *constraint on the performance:* The objective of policy is to ensure, with or without the knowledge or cooperation of the entity being governed, that the policy administrator's intent is carried out with respect to whether or not the specified policy governed action takes place;

[7] While it is true that, over time, norms can be formalized into laws, policies are explicit and formal at the outset by their very nature.

- *machine-executable action:* In addition to purely machine-executable actions, we include situations where a person is responsible for completing an action and then somehow signaling that fact to the machine;
- *subject:* The subject is either a human being or a hardware or software component, or a group of such entities;
- *situation:* Policy applicability may be determined by a variety of preconditions and contextual factors.

16.2.2 Distinguishing Policy from Related Concepts

It is evident that not every constraint in an agent system should be managed as an element of policy. Nor should policy in the sense we are discussing it here be confused with other related concepts. For example, the term *policy* is often used to describe what we will call a "Big P" policy, referring to the sorts of high-level declarations of objectives or preferences that one finds in discussions of strategic policy, public policy, or foreign policy. While it is true that every policy of the sort we are concerned with (call them "little p") is motivated by some higher level objective, "Big P" policies comprise a diversity of elements, some of which involve real world considerations that go far beyond distributed computing issues. Resolving the ambiguities and contradictions of complex and "soft" goals, guidelines, and tradeoffs at the "Big P" level is more the stuff of human deliberation and automated planning than of policy management frameworks which are best suited to analysis and implementation of well-understood constraints *after* the difficult preliminary framing has been done.[8]

Policy management also should not be confused with planning or workflow management, which are related but separate functions. Planning mechanisms are generally *deliberative* (i.e., they reason deeply and actively about activities in support of complex goals) whereas policy mechanisms tend to be *reactive* (i.e., concerned with simple actions triggered by some environmental event) [16.43]. Whereas plans are a unified roadmap for accomplishing some coherent set of objectives, bodies of policy collected to govern some sphere of activity are made up of diverse constraints imposed by multiple potentially—disjoint stakeholders and enforced by mechanisms that are more or less independent from the ones directly involved in planning. Plans tend to be strategic and comprehensive, while policies, in our sense, are by nature tactical and piecemeal. In short, we might say that while policies constitute the "rules of the road"—providing the stop signs, speed limits, and lane markers that serve to coordinate traffic and minimize mishaps—they are not sufficient to address the problem of "route planning."[9]

[8] The relationship between policy at human and computational levels is a subject we are currently investigating.

[9] For an example of how planning and policy management capabilities can complement on another, see [16.113]. Planning can also be used to help assure successful execution of obligation policies.

Policies should not be mistaken for business rules, for while motivations for business rules sometimes overlap with those for policy-based approaches, these two different attempts to enforce regularities on complex systems have usually maintained a different focus. In a manner similar to the world of policies, we can distinguish between "Big B" and "little b" business rules. A "Big B" business rule "pertains to any of the constraints that apply to the behavior of people in the enterprise, from restrictions on smoking to procedures for filling out a purchase order" [16.58]. On the other hand, "little b" business rules pertain "to the facts which are recorded as data and constraints on changes to the values of those facts. That is, the concern is what data may or may not be recorded in the information system" [16.58]. Like "Big P" policies, "Big B" business rules have a much broader scope than "little p" policies. The "little b" rules, on the other hand, are certainly narrower than "little p" policies, to the extent that the former are restricted to governing the kinds of actions that can be performed on a particular instance of a business database rather than to a broader concept of action in general.

Finally, it should be realized that unwanted circumstances cannot be prevented, nor required events be made to happen, by policy management mechanisms alone. A variety of potential failures must be considered and counteracted in the design of safe and effective agent systems, including extreme events; hardware failure; human error; incorrect system design, specification, or implementation; and inconsistency, redundancy, inaccuracy, or incompleteness of agent knowledge and system information [16.43].

16.2.3 Types of Policy

Drawing on their long history of policy research, Sloman et al.[1.34] define the two major types of policy, *authorizations* and *obligations:*

- "A positive authorization policy defines the actions that a subject is permitted to perform on a target. A negative authorization policy specifies the actions that a subject is forbidden to perform on a target".
- "Obligation policies specify the action that a subject must perform on a set of target objects when an event occurs. Obligation policies are always triggered by events, since the subject[10] must know when to perform the specified action".[11]

[10] In the KAoS policy management framework (see Sect. 16.3), a type of enforcer called an *enabler* can be defined to assist subjects in fulfilling obligations, thus reducing, or ideally eliminating, the need for the agent itself to fully understand the policy and to know when and how to undertake its responsibilities [16.16]. Enablers can also be defined for some types of authorization policies.

[11] Some systems differentiate a second class of obligations that requires a given desired state to be continuously maintained by an unspecified action (e.g., Agent A must maintain at least 10 widgets in the bin) in contrast to normal obligations that require a specific action to be performed in response to a trigger (e.g., IF

16.2.4 Autonomy and Policy

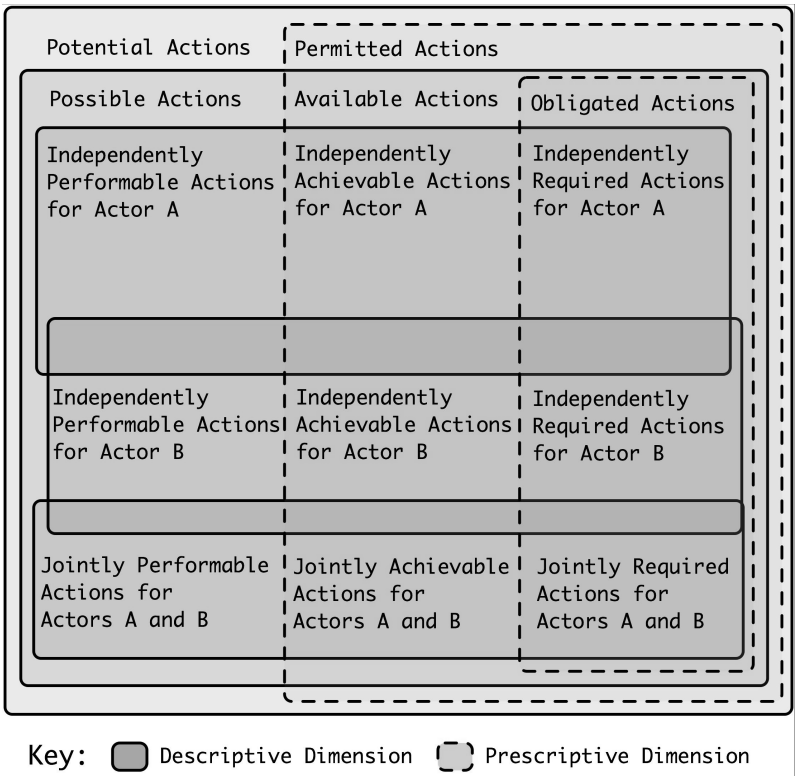

Fig. 16.7. Dimensions of autonomy

Some important dimensions relating to autonomy can be straightforwardly characterized by reference to Fig. 16.7 [12] Note that there are two basic dimensions:

the number of widgets <= 10 THEN Agent must fill the bin with widgets). For example, Pynadath and Tambe [16.76] distinguish between four classes of *safety constraints:* forbidden actions, forbidden states, required actions, and required states. In KAoS (see Sect. 16.3), forbidden actions correspond to negative authorization policies, while required actions and states map to positive obligation policies. Since many states of the world are outside of system control and cannot be forbidden *a priori*, they can best be handled by representing a forbidden state (ideally with some safety margin) as a trigger to a positive obligation policy that requires the agent to achieve a permissible state.

[12] These dimensions are explained more fully in [16.112]. Note that in this section we emphasize those dimensions that are most pertinent to our discussion of adjustable autonomy; see elsewhere for examples of other possible dimensions (e.g., self-impositions, norms). We can make a rough comparison between

– a descriptive dimension-corresponding to the sense of autonomy as self-sufficiency-that stretches horizontally to describe the actions an actor in a given context is capable of performing; and
– a prescriptive dimension-corresponding to the second sense of autonomy as self-directedness-running vertically to describe the actions an actor in a given context is allowed to perform or which it must perform by virtue of policy constraints in force.

The outermost rectangle, labeled potential actions, represents the set of all actions across all situations defined in some ontology under current consideration. [13] Note that there is no requirement that all actions that an actor may take be represented in the ontology; only those which are of consequence for policy representation and reasoning need be included. The rectangle labeled possible actions represents the set of potential actions whose performance by one or more actors is deemed plausible in a given situation [16.116, 16.117]. [14] Note that the definition of possibilities is strongly related to the concept of affordances [16.118, 16.119], in that it relates the features of the situation to classes of actors capable of exploiting these features in the performance of actions. [15] Of these possible actions, only certain ones will be deemed performable for a given actor [16](e.g., Actor A) in a given situation. Capability, i.e., the power that makes an action performable, is a function of the abilities (e.g., knowledge, capacities, skills) and conditions (e.g., ready-to-hand resources) necessary for an actor to successfully undertake some action in a given context. Certain actions may be independently performable by either Actor A or B; other actions can be independently performed by either one

some of these dimensions and the aspects of autonomy described by Falcone and Castelfranchi [16.37]. Environmental autonomy can be expressed in terms of the possible actions available to the agent-the more the behavior is wholly deterministic in the presence of a fixed set of environmental inputs, the smaller the range of possible actions available to the agent. The aspect of self-sufficiency in social autonomy relates to the ranges of what can be achieved independently vs. in concert with others; deontic autonomy corresponds to the range of permissions and obligations that govern the agent's choice among actions.

[13] The term ontology is borrowed from the philosophical literature, where it describes a theory of what exists. Such an account would typically include terms and definitions only for the very basic and necessary categories of existence. However, the common usage of ontology in the knowledge representation community is as a vocabulary of representational terms and their definitions at any level of generality. A computational system's "ontology" defines what exists for the program-in other words, what can be represented by it.

[14] The evaluation of possibility admits varying degrees of confidence-for example, one can distinguish mere plausibility of an action from a more studied feasibility. These nuances of possibility are not discussed in this chapter.

[15] As expressed by Norman: "Affordances reflect the possible relationships among actors and objects: they are properties of the world" [16.119].

[16] For discussion purposes, we use the term actor to refer to either a biological entity (e.g., human, animal) or an artificial agent (e.g., software agent, robotic agent).

or the other uniquely. [17] Yet other actions are jointly performable by a set of actors. [18]

Along the prescriptive dimension, declarative policies may specify various permissions and obligations [16.34]. An actor is free to the extent that its actions are not limited by permissions or obligations. Authorities may impose or remove involuntary policy constraints on the actions of actors. Alternatively, actors may voluntarily enter into agreements that mutually bind them to some set of policies for the duration of the agreement. The effectivity of an individual policy specifies when it is in or out of force. The set of permitted actions is determined by authorization policies that specify which actions an actor or set of actors is allowed (positive authorizations or A+ policies) or not allowed (negative authorizations or A- policies) to perform in a given context. [19] The intersection of what is possible and what is permitted delimits the set of available actions. Of those actions that are available to a given actor or set of actors, some subset may be judged to be independently achievable in the current context. Some actions, on the other hand, would be judged to be only jointly achievable.

Finally, the set of obligated actions is determined by obligation policies that specify actions that an actor or set of actors is required to perform (positive obligations or O+ policies) or for which such a requirement is waived (negative obligations or O- policies). Jointly obligated actions are those that two or more actors are explicitly required to perform.

A major challenge in the design of intelligent systems is to ensure that the degree of autonomy is continuously and transparently adjusted in order to meet whatever performance expectations have been imposed by the system designer and the humans and agents with which the system interacts [16.110, 16.112, 16.120]. We note that is not the case that "more" autonomy is always better: [20] as with a child left unsupervised in city streets during rush hour, an unsophisticated actor insufficiently monitored and recklessly endowed with unbounded freedom may pose a danger both to itself and to others. On the

[17] Note that Figure 16.7 does not show every possible configuration of the dimensions, but rather exemplifies a particular set of relations holding for the actions of a particular set of actors in a given situation. For example, although we show A and B sharing the same set of possible actions, this need not always be the case. Also, note that the range of jointly achievable actions has overlap only with Actor B and not Actor A.

[18] Authority relationships may be, at the one extreme, static and fixed in advance and, at the other, determined by negotiation and persuasion as the course of action unfolds.

[19] We note that some permissions (e.g., network bandwidth reservations) involve allocation of finite and/or consumable resources, whereas others do not (e.g., access control permissions). We note that obligations typically require allocation of finite abilities and resources; when obligations are no longer in effect, these abilities and resources may become free for other purposes.

[20] In fact, the multidimensional nature of autonomy argues against even the effort of mapping the concept of "more" and "less" to a single continuum.

other hand, a capable actor shackled with too many constraints will never realize its full potential.

Thus, a primary purpose of adjustable autonomy is to maintain the system being governed at a sweet spot between convenience (i.e., being able to delegate every bit of an actor's work to the system) and comfort (i.e., the desire to not delegate to the system what it can't be trusted to perform adequately). [21]

The coupling of autonomy with policy mechanisms gives the agent maximum freedom for local adaptation to unforeseen problems and opportunities while assuring humans that agent behavior will be kept within desired bounds. If successful, adjustable autonomy mechanisms give the added bonus of assuring that the definition of these bounds can be appropriately responsive to unexpected circumstances.

In principle, the actual adjustment of an agent's level of autonomy could be initiated either by a human being, the agent, or some other software component. [22] To the extent we can adjust agent autonomy with reasonable dynamism (ideally allowing handoffs of control among team members to occur anytime) and with a sufficiently fine-grained range of levels, teamwork mechanisms can flexibly re-negotiate roles and tasks among humans and agents as needed when new opportunities arise or when breakdowns occur. Such adjustments can also be anticipatory when agents are capable of predicting the relevant events [16.104, 16.105]. Research in adaptive function allocation-the dynamic assignment of tasks among humans and machines-provides some useful lessons for implementations of adjustable autonomy in intelligent systems [16.105].

When evaluating options for adaptively reallocating tasks among team members, it must be remembered that dynamic role adjustment comes at a cost-in both computational and human terms. Measures of expected util-

[21] We note that reluctance to delegate can also be due to other reasons. For example, some kinds of work may be enjoyable to people-such as skilled drivers who may prefer a manual to an automatic transmission.

[22] Cohen and Fleming [16.105] draw a line between those approaches in which the agent itself wholly determines the mode of interaction with human beings (mixedinitiative) and those where this determination is imposed externally (adjustable autonomy). Additionally, mixed-initiative systems are considered by Cohen and Fleming to generally consist of a single user and a single agent. However, it is clear that these two approaches are not mutually exclusive and that, in an ideal world, agents would be capable of both reasoning about when and how to initiate interaction with human beings and subjecting themselves to the external direction of whatever set of explicit authorization and obligation policies were currently in force to govern that interaction. Additionally, there is no reason to limit the notion of "mixed initiative" systems to the single agent-single human case. Hence, we prefer to think of mixed-initiative systems as being systems that are capable of making context-sensitive adjustments to their level of social autonomy (i.e., their level or mode of engagement with human beings), whether a given adjustment is made as a result of reasoning internal to the agent or due to externally imposed policy-based constraints.

ity can be used to evaluate the tradeoffs involved in potentially interrupting the ongoing activities of agents and humans in such situations to communicate, coordinate, and reallocate responsibilities [16.105, 16.108, 16.109]. It is also important to note that the need for adjustments may cascade in complex fashion: interaction may be spread across many potentially distributed agents and humans who act in multiply connected interaction loops. For this reason, adjustable autonomy may involve not merely a simple shift in roles among a human-agent pair, but rather the distribution of dynamic demands across many coordinated actors. [23] Defining explicit policies for the transfer of control among team members and for the resultant modifications required to coordination constraints can prove useful in managing such complexity [16.111]. Whereas goal adoption and the commitment to join and interact in a prescribed manner with a team sometimes occurred as part of a single act in early teamwork formulations, researchers are increasingly realizing the advantages of allowing the acts of goal adoption, commitment to work jointly with a team, and the choice of specific task execution strategies to be handled with some degree of independence [16.7, 16.110].

16.2.5 Benefits of Policy Management

A policy-based approach has many benefits:

Explicit license for autonomous behavior. Policy representations that allow the description of entities and actions at abstract levels (e.g., ontologies) can beneficially underspecify the constraints of policy, giving human stakeholders as much leeway as they require to shape the limits of agent behavior across an arbitrarily large scope of action, while leaving every unmentioned detail completely in the hands of the agents that are closest to the problem. Thus, the coupling of policy with autonomy enables human organizations to *think globally while acting locally.* In short, rather than mistakenly thinking of policy only as a restrictive nuisance, we might more productively think of it as the explicit license by which agents are authorized to make specific decisions and adaptations autonomously in response to novel problems and opportunities as they arise—without violating the constraints imposed by those who are responsible for their behavior.

Reusability. Policies encode sets of useful constraints on agent or component behavior, packaging them in a form in which they can easily be reused as occasions require. By reusing policies when they apply, we reap the lessons learned from previous analysis and experience while saving the time it would

[23] As Hancock and Scallen [16.105] rightfully observe, the problem of adaptive function allocation is not merely one of technical elegance. Economic factors (e.g., can the task be more inexpensively performed by humans, by agents, or by some combination?), political and cultural factors (e.g., is it acceptable for agents to perform tasks traditionally assigned to humans?), or personal and moral factors (e.g., is a given task enjoyable and challenging vs. boring and mind-numbing for the human?) are also essential considerations.

have taken to reinvent them from scratch. Policy libraries can package sets of policies that have been pre-approved for particular situations. For example, military applications may have different policy sets defined that come into play for various levels of threat conditions.

Efficiency. In addition to lightening the application developers' workload, well-defined policy management mechanisms can sometimes increase runtime efficiency [16.76]. For example, to the extent that policy conflict resolution can be performed offline in advance, and policies can be converted to an efficient runtime representation, overall performance can be increased [16.16, 16.117].

Extensibility. A well designed policy management capability provides a layer of basic representations and services that can be straightforwardly extended to diverse and evolving platforms and to sets of operational capabilities that are often subject to rapid rates of technology refresh. Ideally, these modifications could be made without extensive manual markup or duplication of information stored elsewhere in the organization.

Context-sensitivity. Explicit policy representation improves the ability of agents, components, and platforms to be responsive to changing conditions without changing their code. In mature policy management systems, such changes to policy can be either made manually through convenient distributed administration capabilities or triggered programmatically by events.

Verifiability. By representing policies in an explicitly declarative form instead of burying them in the implementation code, we can better support important types of policy analysis [16.43]. First—and this is absolutely critical for security policies—we can externally validate whether or not the policies are sufficient for the application's tasks, and we can bring both automated theorem provers and human expertise to this task. Second, there are methods to ensure that agent behavior which follows the policy will also satisfy many of the important properties of reactive systems: liveness, recurrence, safety invariants, and so forth.

Support for simple as well as sophisticated agents. By putting the burden for policy analysis and enforcement on the infrastructure, rather than having to build such knowledge into each of the agents themselves, we ensure that all agents operate within the bounds of policy constraints [16.15]. In this way, even one agent shall not be lost due to policy violations, no matter how simple or sophisticated the agent's design, and the task of agent developers is thereby reduced in complexity [16.48].

Protection from poorly-designed, buggy, or malicious agents. Intelligent systems functioning in complex environments cannot rely on design-time techniques to completely eliminate the possibility of unwanted events occurring during operations.[24] Moreover, even if it could be guaranteed that agents designed by a given group would always function correctly, the fact

[24] As Fox and Das [16.43] wisely observe, "the nature of a hazard will frequently be unknown until it actually arrives. In some circumstances, ensuring that a system reliably does what the designers intended—and only what they intended—may be exactly the wrong thing to do!"

is that, as long as reliance on open systems continues to increase, the possibility of buggy or malicious agents designed by others cannot be completely ignored. Various forms of policy-based barriers that can control the actions of such agents through monitoring, analysis, inference, adjustable autonomy, and enforcement methods that are infrastructure-based and independent of the agents' own reasoning, appear to be the most effective ways to reduce the risk of these serious problems [16.59].

Reasoning about agent behavior. As permitted by disclosure policies [16.83], sophisticated agents can reason about the implications of the policies that govern their behavior and the behavior of other agents. To the extent that behavior can be predicted from policy, making accurate and consistent models of agents becomes more feasible.

16.2.6 Applications of Policy Using KAoS and Nomads

At the Institute for Human and Machine Cognition (IHMC), we have developed KAoS and Nomads to support a wide range of policy and domain services. KAoS a collection of componentized policy and domain management services compatible with several popular agent frameworks, including Nomads, the DARPA CoABS Grid, the DARPA ALP/Ultra*Log Cougaar framework (http://www.cougaar.net), CORBA (http://www.omg.org), Voyager (http://www.recursionsw.com/osi.asp), Brahms (www.agentisolutions.com), TRIPS [16.2, 16.3, 16.38], and SFX (http://crasar.eng.usf.edu /research/publications.htm). While initially oriented to the dynamic and complex requirements of software agent applications, KAoS services are also being adapted to general-purpose grid computing (http://www.gridforum.org) and Web Services (http://www.w3.org/2002/ws/) environments as well. A comparison between KAoS, Rei, and Ponder for policy specification, representation, reasoning, and enforcement is given in [16.121]. More complete descriptions of KAoS and Nomads can be found in [16.112, 16.113].

To help motivate a later discussion of different kinds of policy, and to give some idea of the wide range of problems to which policy-based approaches can be applied, we briefly describe some applications.

The DARPA CoABS-sponsored Coalition Operations Experiment (CoAX) (http://www.aiai.ed.ac.uk/project/coax/) is a large international cooperation that models military coalition operations and implements agent- based systems to mirror coalition structures, policies, and doctrines. CoAX aims to show that the agent-based computing paradigm offers a promising new approach to dealing with issues such as the interoperability of new and legacy systems, the implicit nature of coalition policies, security, and recovery from attack, system failure, or service withdrawal [16.4]. The most recent CoAX-related work also investigates issues in composition of semantic web services consistent with negotiated policy constraints [16.114]. KAoS provides mechanisms for overall management of coalition organizational structures represented as domains and policies, while Nomads provides strong mobility,

resource management, and protection from denial-of-service attacks to untrusted agents that run in its environment.

Within the DARPA Ultra*Log program (http://www.ultralog.net), we are collaborating with CougaarSoft to extend and apply KAoS policy and domain services to assure the scalability, robustness, and survivability of logistics functionality in the face of information warfare attacks or severely constrained or compromised computing and network resources. In agent societies of over 1,000 agents and hundreds of policies, dynamic policy updates can be committed and distributed across multiple hosts in a matter of seconds, and responses to policy authorization queries average less than 1 ms [16.122].

As part of the Army Research Lab Advanced Decision Architectures Consortium, we have been investigating the use of KAoS and Nomads technologies to enable soldiers in the field to use agents from handheld devices to perform functions such as dynamically tasking sensors and customizing information retrieval. Suri has developed an agile computing platform that provides a foundation for this work [16.90, 16.91, 16.92, 16.94]. We have also commenced an investigation of requirements for policy-based information access and analysis within intelligence applications.

An application focused more on the social aspects of agent policy is within the NASA Cross-Enterprise and Intelligent Systems Programs [16.23], where we are investigating the integration of Brahms, an agent-based design toolkit that can be used to model and simulate realistic work situations in space, with KAoS policy-based models and Nomads's strong mobility and resource control capabilities to drive human-robotic teamwork and adjustable autonomy for highly-interactive autonomous systems, such as the Personal Satellite Assistant (PSA). The PSA is a softball-sized flying robot that is being developed to operate onboard spacecraft in pressurized micro-gravity environments [16.44]. The same approach has also being generalized for use in mobile robots for planetary surface exploration [16.85]. The Office of Naval Research (ONR) is supporting research to extend this work on effective human-agent interaction to unmanned vehicles and other autonomous systems that involve close, continuous interaction with people. As one part of this research, IHMC and the University of South Florida are developing a new robotic platform with carangiform (fish-like) locomotion, specialized robotic behaviors for humanitarian demining, human-agent teamwork, agile computing, and mixed-initiative human control.

We are investigating issues in adjustable autonomy and mixed-initiative behavior for software assistants under funding from the DARPA EPCA (CALO) program [16.112, 16.120]. Under funding from DARPA's Augmented Cognition Program, we are also taking the challenge of effective human-agent interaction one step further as we investigate whether a general policy-based approach to the development of cognitive prostheses can be formulated, in which human-agent teaming could be so natural and transparent that robotic

and software agents could appear to function as direct extensions of human cognitive, kinetic, and sensory capabilities (see Sect. 16.4.2).

16.3 Technical Aspects of Agent Acceptability

Norman suggests that the technical considerations include such things as ensuring robustness against technical failures, guarding against error and maliciousness, and protecting privacy [16.72]. We touch on each of these considerations in some way in this section. Later on in the chapter (Sect. 16.4.1), we present examples of policies relating to social aspects of agent behavior. Admittedly the distinction between the two kinds of examples is not always clearcut.

Examples of the kinds of basic infrastructure that will be required to support the technical aspects of agent acceptability are becoming more available. Designed from the ground up to exploit next-generation Internet and Web-Services capabilities, grid-based approaches, for example, aim to provide a universal source of dynamically pluggable, pervasive, and dependable computing power, while guaranteeing levels of security and quality of service that will make new classes of applications possible ([16.42]; http://www.gridforum.org). By the time these sorts of approaches become mainstream for large-scale applications, they will also have migrated to ad hoc local networks of very small devices [16.45, 16.92].

This being said, however, we must go far beyond these current efforts to enable the vision of long-lived agent communities performing critical tasks (Fig. 16.8). Current infrastructure implementations typically provide only very simple forms of resource guarantees and no incentives for agents and other components to look beyond their own selfish interests. At a minimum, future infrastructure must go beyond the bare essentials to provide pervasive *life support services* (relying on mechanisms such as orthogonal persistence and strong mobility [16.88, 16.89]) that help ensure the survival of agents that are designed to live for many years. Beyond the basics of individual agent protection, long-lived agent communities will depend on *legal services*, based on explicit policies, to ensure that individual and societal rights and obligations are monitored and enforced. Benevolent *social services* might also be provided to proactively adjust autonomy to avoid problems and help agents fulfill their obligations [16.112]. Although some of these elements exist in embryonic stage within specific agent systems, their scope and effectiveness has been limited by the lack of underlying support at both platform and application levels.

Concern	Service Level	Benefit
Welfare	Social Services	Get help when needed
Justice	Legal Services	Get what you deserve
Environmental protection	Life Support Services	Get enough to survive
Looking out for #1	Bare Essentials	Get what you can take

Fig. 16.8. Required elements of future infrastructure for software agents

16.3.1 Examples of Policy Types Relating to Technical Aspects of Agent Acceptability

To better describe the nature of policy as it relates to the technical aspects of agent acceptability, we now discuss several examples. These examples are intended not to be comprehensive but illustrative. Some of them are related to actual policies that we have used in various applications of KAoS; others reflect cases we have anticipated but not yet implemented.

For clarity, we will present example policies in ordinary English rather than in OWL (Web Ontology Language, used to represent KAoS policies). For brevity, the policies will be presented in an incomplete, abbreviated form. Each example is preceded by A+, A-, O+, or O- to indicate whether it is, respectively, a positive authorization, a negative authorization, a positive obligation, or a negative obligation. Note that although we present many of the policies in "IF ... THEN" form for convenient exposition, such conditional information is actually represented in KAoS in the form of OWL property restrictions on action classes rather than in rules. We will look at six categories of technical policy: authentication, data and resource access and protection, communication, resource control, monitoring and response, and mobility.

Authentication policies.

O+: *IF KPAT is launched*
THEN that instance of KPAT is required to successfully complete a strong authentication process within time T
PRECEDENCE: A-: no one can use this instance of KPAT
ELSE O+: this instance of KPAT must terminate.

In this example, which is typical of some of the policies developed within our DARPA Ultra*Log research, the policy assures that strong authentication will be performed each time an effort is made to launch KPAT. Strong

authentication is an abstract action that can represent any number of more specific strong authentication methods in the ontology that are available to the system. The authentication might be performed by KPAT itself or delegated to an enabler. One could argue that this policy should be hard wired into the code rather than represented explicitly. That, however, would reduce flexibility in ways that may not be desirable. For example, KPAT administrators at times may want to take this policy out of force in an emergency situation.

In OWL, we represent the precedence conditions as one or more policies. In this case, a negative authorization policy forbids any use of KPAT until the conditions of the obligation policy are fulfilled. Roles are represented straightforwardly as merely one kind of domain or group in which human or agent actors belong. It is recommended to use time, or some more general state indicator, as one of the conditions of obligation fulfillment in order to minimize the risk of the agent getting "stuck" indefinitely. Consequences of non-fulfillment of the obligation (the "ELSE" clause) are also represented as policies. In this case, KPAT is obliged to terminate if the obligation is not successfully fulfilled. In subsequent examples, we will not always list the precedence, conditions of fulfillment, or consequences of non-fulfillment explicitly.

A-: *A user is forbidden from taking any action with account A*
IF the user has login_failure_count >= n and time since failure <= T

This negative authorization policy, again representative of our Ultra*Log work, deals with authentication failure. After a given number of login failures, the user is locked out of the account until some period of time elapses.

O+: *IF the space station crew member has issued a voice command*
THEN the Personal Satellite Assistant (PSA) is required to authenticate the crew member's voice within time T
PRECEDENCE: A-: PSA is forbidden to perform the action corresponding to the command
ELSE O+: PSA notifies crew member appropriately

This example is drawn from our NASA human-agent teamwork research. Since authorization for some PSA action may depend on who commanded it, authentication of the crewmember's voice is required before the action is performed. The "ELSE" clause embeds a notification policy.

Data and resource access and protection policies.

A-:*Agent X is forbidden from saving data that is unsigned and/or unencrypted*

This data protection policy example specifies that agent X must sign and encrypt all data that it saves. In our work with Ultra*Log, the encryption would be performed by an enabler. In other words, each time X saves data, the policy is enforced through the enabler transparently doing the proper sort of encryption on its behalf.

A-: *All actors in Role R are forbidden from performing any action on servlet S*

This resource access policy prevents any unauthorized use of Java servlet S by actors (i.e., agents or humans) who are in role R. As in a previous example, the power of abstract specification in the ontology is highlighted: note that this policy can be specified without having to know in advance the particular actions that can be performed on the servlet.[25] Additional Ultra*Log examples of this kind include policies governing actions such as Java JAR file verification, limiting access to private keys, and predicate-based access restriction to blackboard information.

A+: *Users in Role CA Administrator are permitted to perform the revoke certificate operation on the CA Service*

Users or agents in a given role or with a given privilege are authorized here to revoke certificates.

An important part of our current investigations on policy-based information access for intelligence applications concerns disclosure policies. These sorts of policies control the kinds of intelligent responses that can be given as part of queries about which policies are relevant to a given user's analysis or decision-making context [16.114]. In a related application, an agent may want to know about the policies of a given domain before it registers to join. Disclosure policies would determine what kind of policy information could be given to that agent without compromising confidentiality. We are drawing on the work of [16.83, 16.122] to develop more complex strategies for policy disclosure and automated trust negotiation in a variety of circumstances.

Communication policies. Communication has proven to be the most important application of policy within our CoAX research [16.4]. Typically, the domains are configured to be in the "tyrannical" mode, blocking communication among different countries, organizations, or functional groups unless otherwise specified. For example, administrators from the fictional country of Arabello decided on the following restrictive default policy for the actors in their domain:

A-: *Agents in the Arabello domain are forbidden from sending messages to any agent outside the Arabello domain*

However administrators from the Arabello contingent wanted to enable the Arabello Intel agent to be able to send a subset of its reports to the coalition. They specified the following policy, which was assigned a higher precedence for policy conflict resolution purposes:

[25] Capability-based access is a term used by Suri to describe an additional level of protection, where all of the details of service implementation are hidden from the client for confidentiality purposes.

A+: *Arabello Intel Agent is permitted to send messages about enemy diesel submarines to any member of the Binni-Coalition domain (sharing messages about any other topic is still forbidden)*

Communication blocking based on message content as illustrated in this example is facilitated by the use of a custom editor within KPAT that allows the administrator to specify the kinds of messages that are to be permitted based on OWL-typing of various message fields [16.90].

A+: *MAD Sensor Agent is permitted to send reports with image resolution X:Y to any member of the Arabello domain*

As part of CoAX, as well as in follow-on Army research, we have also addressed requirements for filtering and transformation of data [16.90, 16.94]. For example the provider of a Magnetic Anomaly Detector (MAD) sensor was willing to share its reports with Arabello, but only on condition that the sensor's signal could be appropriately downgraded in order to prevent Arabello from knowing the full extent of the sensor's capabilities. The policy enforcer-enabler used in this application could be configured by policy to allow three different types of data transformation: a) changes in image resolution, b) changes in frame rate and c) introduction of time lags to prevent transmission of a real time video feed.

Many other types of policy-based transformations could be envisioned for sensor data feeds. A policy enforcer-enabler could, for instance, be implemented to hide sensitive targets or classified infrastructures from the image. This would be used to prevent the release of unnecessary details to the requesting agent by blurring or editing the image appropriately. Another example is an agent that reduces the precision of coordinate values embedded in message content. More generally, such filtering and transformation techniques can be used for sources and methods protection, and as part of the management of information pedigrees and digital rights protection.

In the Ultra*Log application, policies are required to block both sending and receiving of certain kinds of messages. The fact that KAoS policies can specify whether the site of enforcement is to be associated with the subject or the target is useful for this purpose: both the sending and the receiving can be blocked at either the subject or the target side as convenience dictates. *Policy templates* developed for Ultra*Log allow users to specify a composite set of multiple policies more simply as if it were a single policy. To take a simple example, the details of blocking of both sending and receiving messages are accomplished through a simple user interface that presents policy specification options in terms of the more general concept of "communication blocking." As additional examples of communication policies, Ultra*Log also requires that administrators be able to specify which cryptographic modes, transport types, and message formats are allowable in a given context. It also requires limits on message size and system resources in message delivery.

Resource control policies. Whereas resource access policies govern whether or not a resource is made available, resource control policies go a step further to control the amount and rate of resource usage (e.g., CPU, memory, network, hard disk, screen space). For example as part of one of the CoAX scenarios the country of Gao requests permission to host one of its agents on a sensor platform. Because its intentions are unclear and it is distrusted, it is required to run on top of the Aroma VM. Because the Aroma VM is Java-compatible, Gao is not aware of this restriction. Later, when Gao's agent launches a denial-of-service attack which floods the network and begins consuming inordinate amounts of CPU and disk resources, the pattern of misuse is noticed by a Guard, which has been previously authorized to automatically lower the resource limits enforced by the Aroma VM in such situations by one or more policies, such as the following:

O+: *IF a Guard notices a pattern of resource misuse by an agent*
THEN that Guard must notify its administrator appropriately
PRECEDENCE A-: The agent is forbidden from using more than 25% of the resource
ELSE A-: The agent is forbidden from using more than 10% of the resource

The policy requires the Guard to notify the administrator, who can determine whether this is a false alarm (in which case the agent's resources can be restored by a new policy setting) or whether this is a real attack (in which case the administrator may choose to further lower A's resource usage). If the effort to notify the administrator fails, the Guard is authorized to reduce resource usage to 10% on its own. In this case, transparently reducing resource usage is better than peremptorily terminating the agent because in the former case the agent will be unaware that it's misuse has been detected.

The requirement for the Guard to be able to act autonomously in making an initial response to the attack is akin to the need for a sprinkler system in a building to go off in the presence of smoke before the fire department arrives. Though there is a risk that the signal may have been a false alarm, it is still far better in most cases to have limited the potential damage through prompt action. Moreover, in the case of a malicious agent that is attacking the network, the administrator may not be able to reconfigure a remote sensor until a provisional limit is placed on network resource usage.

A+: *Team A is authorized to use 50% of the CPU*

In order to guarantee a certain quality of service to other agents, Team A is limited to 50% in the amount of CPU resources it is authorized to use. In this example, however, note that the policy says nothing about how the CPU resources should be allocated among members of Team A, so internal resource allocation is left to the particular algorithm used by the enforcer performing this task.

Monitoring and response policies. It may sometimes be desirable to represent obligations, for the system to perform specific monitoring and response actions as policy:

O+: *IF an authorization failure event occurs*
THEN the authorization mechanism must record the pertinent data in the system log
PRECEDENCE A-: the authorization mechanism is forbidden to perform any other action
ELSE O+: the authorization mechanism must notify the administrator appropriately

In this example, the authorization mechanism is required to record pertinent data in the system log if an authorization failure event occurs. In another example from Ultra*Log:

O+: *IF there is a new defense posture*
THEN the policy applicability condition monitor must deploy the M&R component group for the new defense posture and deactivate the M&R component for the previous defense posture
PRECEDENCE A-: the policy applicability condition monitor is forbidden to perform any other action
ELSE O+: the policy applicability condition monitor must notify the administrator appropriately

This policy requires a new set of monitoring and response components to be activated when the defense posture changes (e.g., a change from threatcon alpha to threatcon bravo).

Mobility policies.

A-: *Agents that are members of the trust domain are forbidden from moving to host H*

This example illustrates how the movement of software agents from one host to another can be controlled by policy in the same way that any other action is governed, provided appropriate enforcement mechanisms are in place.

In a more complex example based on research by Knoll et al. [16.57], the trust level of a mobile software agent is determined in part by where it has traveled in the past (i.e., there is greater or lesser possibility that it may have been tampered with by a malicious host). The trust level, in turn, is used to limit the permissions of the agent in the future:

A-: *Agents are forbidden from performing sensitive action X*
if their trust level <= threshold

The following example, pertinent to our NASA work on the PSA, obligates the PSA to move away from danger:

O+: *IF a situation dangerous to a PSA is present in some location THEN the PSA must move out of that location*[26]

16.4 Social Aspects of Agent Acceptability

Norman suggests that the social aspects of agent acceptability include things such as providing reassurance that everything is working according to plan, providing an understandable and controllable level of feedback about agent's intentions and actions, and accurately conveying the agent's capabilities and limitations [16.72]. In short, human beings must be informed enough to be able to easily step in and help when the situation becomes more than the agents can handle, and agents on their part must be made more competent in conveying the appropriate information to humans and acting in partnership with them. Speaking of the central problems of conventional automation, Norman writes:

> "The problem ... is that automation is at an intermediate level of intelligence, powerful enough to take over control that used to be done by people, but not powerful enough to handle all abnormalities. Moreover, its level is insufficient to provide the continual, appropriate feedback that occurs naturally among human operators. To solve this problem, the automation should either be made less intelligent or more so, but the current level is quite inappropriate ... . Problems result from inappropriate application, not overautomation" [16.70].

Teamwork has become the most widely accepted metaphor for describing the nature of cooperation in multi-agent systems. Whereas early research on agent teamwork focused mainly on agent-agent interaction [16.28, 16.115], teamwork principles are now being formulated in the context of human-agent interaction [16.11, 16.14]. Unlike autonomous systems, designed primarily to take humans out of the loop, many new efforts are specifically motivated by the need to support close continuous multimodal human-agent interaction [16.22, 16.27, 16.53, 16.61, 16.73, 16.114].

The KAoS policy-based teamwork model defines what constitutes a team, and the nature of many of its collaborative activities. Elsewhere, we have outlined a preliminary perspective on the basic principles and pitfalls of adjustable autonomy and human-centered teamwork gleaned from the literature [16.14]. The set of policies we are designing for human-robotic interaction goes beyond the traditional policy concerns about security and safety in significant ways. As one example, consider how policy can be used to ensure effective communication among team members. Previous research on generic

[26] Consistent with Asimov's laws; however, the PSA might be obliged by a higher-level policy to stay if its presence was needed to help a human being.

teamwork models has explored this issue to a limited degree within the context of communication required to form, maintain, and abandon joint goals. However, more research is needed to address the complexities of maintaining mutual awareness in human-agent, as opposed to agent-agent, interaction.

With previous research in agent teamwork, we share the assumption that, to the extent possible, teamwork knowledge should be modeled explicitly and separately from the problem-solving domain knowledge. Policies for agent safety and security, as well as context- and culturally-sensitive teamwork behavior, can be represented as KAoS policies that enable many aspects of the nature and timing of the agent's interaction with people to be appropriate, without requiring each agent to individually encode that knowledge. Agent designers can concentrate on developing unique agent capabilities, while assuming that many of the basic rules of effective human-agent coordination will be built into the environment as part of the policy infrastructure.

16.4.1 Examples of Policy Types Relating to Social Aspects of Agent Acceptability

In contrast to the examples of technical policies in Section 16.3.1 above, our work to begin encoding these social issues in policy is relatively recent and is likely to evolve considerably in the near future. Some of this will require the resolution of difficult research issues; we are beginning implementation with those policies that are most straightforward, and will then continue to progress incrementally to more complex ones. We will give examples from six categories of policy: organization, notification, conversation, nonverbal expression, collaboration, and adjustable autonomy.

Organization policies. Some policy management systems, in part as an artifact of their mode of policy representation, require many or all of what we call organization policies to be represented as "meta level," "higher order," or some other sort of special policy. In KAoS, many of these can be specified uniformly, in the same way that other kinds of policies are represented.

A+:*Individuals of the class Domain Manager are permitted to approve policies*

The KAoS actor ontology distinguishes between people and various classes of agents. Most agents can only perform *ordinary actions*, however various components that are part of the infrastructure (e.g., domain manager, guard), as well as authorized human users, may variously be permitted or obligated to perform *policy actions,* such as policy approval and enforcement.

A+: *Any person in the Manager Role is permitted to authorize check payment IF the same person is not also the check issuer*

The example specifies a dynamic separation of duty, where the issuer of the check is not allowed to also be the one who authorizes payment on that check.

A-: *An agent is forbidden to register to domain D IF it is already registered to any individual of the class domain*

A-: *An agent is forbidden to register to any individuals of the class domain IF it is already registered to domain D*

The pair of policies above specifies that an agent cannot simultaneously be registered as a member of both the domain D and some other domain.

Notification Policies. Building on the work of [16.81, 16.97], we are developing KAoS notification policies in the context of our NASA applications. The vision of future human-agent interaction is that of loosely coordinated groups of humans and agents. As capabilities and opportunities for autonomous operation grow in the future, agents will perform their tasks for increasingly long periods of time with only intermittent supervision. Most of the time routine operation is managed by the agents while the human crews perform other tasks. Occasionally, however, when unexpected problems or novel opportunities arise, humans must assist the agents. Because of the loose nature of these groups, such communication and collaboration must proceed asynchronously and in a mixed-initiative manner. Humans must quickly come up to speed on situations with which they may have had little involvement for hours or days. Then, they must cooperate effectively and naturally with the agents as true team members.[27] Hence the challenge of managing notification and situation awareness for the crewmembers.

Various ontologies supporting notification (e.g., basic concepts for categories of events, roles, notifications, latency, focus of attention, and presence) form the foundation of this work. In conjunction with these ontologies, notification policies and their parameter settings are created, as in the example below:

O+: *IF new notification = true AND utility >= notifyThreshold AND utility < doItThreshold*
THEN notify the space station crew member appropriately

Human attention is a scarce resource. When an important event is signaled, the utility of various alternatives (e.g., notify the crew member, perform some required action without interrupting the person, or do nothing) is evaluated. If a notification is required and the current task is well-defined, the KAoS-Brahms infrastructure will take into account the task and other contextual factors to perform the notification in a manner that is context-sensitive to modality, urgency, and the location of the human being. Because

[27] Actually, this vision points to two major opportunities for policy-based help: 1. the use of policy to assure that unsupervised autonomous agent behavior is kept within safe and secure limits of prescribed bounds, even in the face of buggy or malicious agent code; and 2. the use of policy to assure effective and natural human-agent team interaction, without individual agents having to be specifically programmed with the knowledge to do so.

such knowledge resides in the infrastructure, rather than as part of the knowledge of each agent, agent development is simplified.

Conversation Policies. Explicit conversation policies simplify the work of both the agent and the agent designer [16.12, 16.13, 16.47, 16.48]. In comparison to unrestricted agent dialogue, conversation policies reduce the agents' inferential burden by limiting the space of alternative conversational productions and parameters that they need to consider, both in generating messages and in interpreting messages received from other agents. Because a significant measure of conversational planning for routine interactions can be encoded in conversation policies offline and in advance, the agents can devote more of their computational power at runtime to other things.

O+: *IF response lag of conversation X > M minutes*
THEN the agent must terminate conversation X

This conversation policy requires the agent to unilaterally terminate a conversation when a lag of M minutes has elapsed in waiting for a response, preventing conversations from staying open indefinitely.

A-:*Agents are forbidden to send a message of any type other than Reply*
IF the type of the conversation is Request-Reply AND the previous message type of the conversation is Request

This conversation policy enforces a sequence of messages in the Request-Reply conversation type, such that a message of type *request* must always be followed by a message of type *reply*.

A+: *Agents are permitted to send TRANSCOM messages with return receipts*

This conversation policy example from the Ultra*Log application, allows the sending of TRANSCOM messages that require return receipts. Note that this policy would be appropriate only in a tyrannical domain that prohibited all messages that were not explicitly permitted.

More complex sorts of policies, dealing, for example, with Clark's concept of common ground [16.24] or improvisational approaches to conversation [16.78], will also be important in effective human-agent interaction. Though such policies go beyond what is possible in the current version of KAoS, we have started to address them as part of a collaboration with Allen et al. [16.2, 16.3] in the near future [16.112, 16.120].

Nonverbal Expression Policies. Where possible, agents usually take advantage of explicit verbal channels for communication in order to reduce the need for relying on current primitive robotic vision and auditory sensing capabilities [16.69]. On the other hand, animals and human beings often rely on visual and auditory signals instead of explicit verbal communication for many aspects of coordinated activity. As part of our work on human-robotic interaction for NASA, the Army, and the Navy, we are developing policies

to govern various nonverbal forms of expression in hardware and software agents. These nonverbal behaviors will be designed to express not only the current state of the agent but, importantly, also to provide rough clues about what it is going to do next. In this way, people can be better enabled to participate with the agent in coordination, support, avoidance, and so forth. In this sense, nonverbal expressions are an important ingredient in enabling human-agent teamwork.

Maes and her colleagues were among the first to explore this possibility of software agents that continuously communicate their internal state to the user via facial expressions (e.g., thinking, working, suggesting, unsure, pleased, and confused) [16.66]. Breazeal and Scassellati [16.18] have taken inspiration from research in child psychology [16.97] to develop robot displays that reflect four basic classes of preverbal social responses: affective (changing facial expressions), exploratory (visual searching, maintaining mutual regard with a human being), protective (turning away head), and regulatory (expressing feedback to gain caregiver attention, cyclical waxing and waning of internal states, habituating, and signalling internal motivation). Books on human etiquette [16.118] contain many descriptions of appropriate behavior in a wide variety of social settings. Finally, in addition to this previous work, we think that behavior displayed among human beings [16.68] and groups of animals will be one of the most fruitful sources of policy for effective nonverbal expression in agents. Our initial study indicates that there are useful agent equivalents for each of Smith's [16.86] ten categories of widespread vertebrate animal cooperation and coordination displays [16.115].

O+: *IF the current task of the PSA is of type uninterruptible*
THEN the PSA must blink a red light until the current task is finished
PRECEDENCE: A-: The PSA is forbidden from performing any tasks but the current one

This policy requires the PSA to blink a red light while it is busy performing an uninterruptible task. During this time, it is also forbidden from performing any task but the current one. Related messages it may want to give with a similar signal might include: "I am unable to make contact with anybody," "Do not attempt to communicate with me (for whatever reason, e.g., 'my line is bugged')." On the positive side, various uses of a green light might signal messages such as: "I am open for calls," "I need to talk to someone," or "May I interject something into this conversation? " Displays in this general interactional category clearly have benefits for coordination among groups of agents by providing information about which agents are (or are not) in a position to interact with others, in what ways, when, and so forth.

O+: *IF a conversation has been initiated with someone*
THEN the PSA must face the one with whom it is conversing, as long as it is in sight, until the conversation has finished

This policy implements a kind of display associated with maintaining a previously established association. This display might be especially useful when the PSA is moving around the room and needs to let someone else know that it is still attending to an ongoing conversation.

O+: *IF the current task of the PSA is to move some distance greater than D*
THEN the PSA must signal its intention to move for S seconds
PRECEDENCE: A-: The PSA is forbidden from executing its move

It's no fun being hit in the head by a flying robot that suddenly decides to go on the move. This policy prevents the PSA from moving until it has first signaled its intention to move for some number of seconds. Besides the pre-move signaling, some kind of signaling could also take place during the move itself. In addition to this movement signaling policy, other policies should be put in place to require the PSA to stay at a safe and comfortable distance from humans, other robotic agents, and space station structures and equipment.

Collaboration Policies.

O+: *IF an agent becomes aware that a team goal has been achieved, or has become unachievable or irrelevant*
THEN the agent must notify the other team members in an appropriate manner
PRECEDENCE: A-: The team member is forbidden from actions that are performed only in order to achieve the former team goal

A similar version of this policy is one of the centerpieces of the classic theory of teamwork originally proposed by Cohen and Levesque [16.28]. Though there is potentially a lot of complexity buried in the machinery that determines whether the condition is true, the policy imperative that results from this condition is relatively simple and can be represented straightforwardly in KAoS. All the foundational ontologies and mechanisms developed to support other kinds of notification policies can also be brought to bear in this context. In this sense, the example can be seen as just a special kind of notification policy.

O+: *IF an agent suspends work on a current task in order to attend to a new higher priority task*
THEN the agent must notify the other actors involved in an appropriate manner
PRECEDENCE: A-: The agent is forbidden from executing its new task

Just as a sudden physical move might surprise other actors unless it is appropriately signaled in advance, so also an unexpected change in current task might be jarring to others unless it is heralded in some fashion. Note that this policy presupposes that additional actors, beyond the team members themselves, that share a joint goal may require notification.

Adjustable Autonomy Policies. Humans and agents may play mutual roles that vary according to the relative degree of initiative appropriate for a given situation (Fig. 16.9). [28] At the one extreme, traditional systems are designed to carry out the explicit commands of humans with no ability to ignore orders (i.e., executive autonomy), generate their own goals (i.e., goal autonomy), or otherwise act independently of environmental stimuli (i.e., environmental autonomy). Such systems cannot, in any significant sense, act; they can only be acted upon. At the other end of the spectrum is an imagined extreme in which agents would control the actions of humans. [29] Between these two extremes is the domain of today's agent systems, with most agents typically playing fixed roles as servants, assistants, associates, or guides. Such autonomous systems are designed with fixed assumptions about what degree of initiative is appropriate for their tasks. They execute their instructions without considering that the optimal level of autonomy may vary by task and over time, or that unforeseen events may prompt a need for either the human being or the agent to take more control. At the limit of this extreme are strong, silent systems with only two modes: fully automatic and fully manual [16.77]. In practice this can lead to situations of human "underload," with the human being having very little to do when things are going along as planned, followed by situations of human "overload," when extreme demands may be placed on the human in the case of agent failure.

Although in practice many mixed-initiative system do not live up to their billing, their design goal is to allow agents to dynamically and flexibly assume a range of roles depending on the task to be performed and on the current situation. Research in adjustable autonomy supports this goal through the development of an understanding about how to ensure that, in a given context, the agents are operating at an optimal boundary between the initiative of the human being and that of the agents. People want to maintain that boundary at a sweet spot in the tradeoff curve that minimizes their need to attend to interaction with the agent while providing them with a sufficiently comfortable level of assurance that nothing will go wrong.

O+: *IF elapsed time since last report > time T*

[28] For a more fine-grained presentation of a continuum of control between humans and machines, see Hancock and Scallen's [16.50] summary of Sheridan's ten-level formulation. Robert Taylor (personal communication) surmises from his experience that ten may be far more levels than are useful in practice. Barber et al. differentiate three kinds of relationships among agents: command-driven (i.e., the agent is fully subordinated to some other agent), true consensus (i.e., decision-making control is shared equally with other agents), and locally autonomous/master (i.e., the agent makes decisions without consulting other agents and may be allowed to command subordinates).

[29] Of course, in real systems, the relative degree of initiative that could be reasonably taken by an agent or human would not be a global property, but rather relative to particular functions that one or the other was currently assuming in some context of joint work.

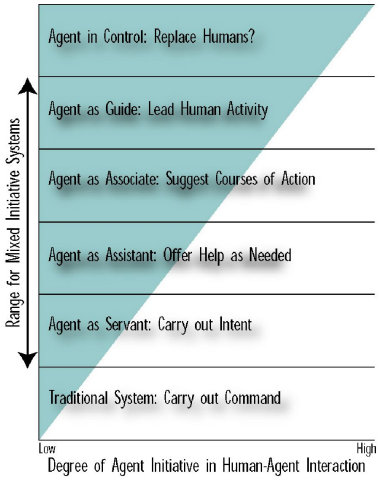

Fig. 16.9. Spectrum of agent roles in human-agent interaction

THEN the agent must notify the human supervisor about its status in the most appropriate manner

This simple policy sets the duration of autonomous operation for an agent, requiring it to notify a human supervisor about its status at predetermined intervals.

A+: *Space station crewmembers in the Trusted Operator Role are permitted to override PSA non-critical negative authorizations*

Sometimes, it is critical for authorized human operators to be able to immediately countermand some negative authorization of an agent (i.e., allowing it to do things which it normally is not authorized to do). While this could be done by modifying the policy in the usual way, it is sometimes more practical to do this directly on a one-time temporary basis by overriding a prohibition. However, overriding certain operations (e.g., flying the PSA into a wall of the space station) may require consent of both the space station commander and an authorized person at mission control.

O+: *IF no crewmember is monitoring the environment in space station module X*

THEN PSA must monitor the environment in module X

PRECEDENCE: A+: PSA is permitted to monitor the environment in module X

Sometimes, the PSA may be required to temporarily take upon itself functions that human crewmembers would normally provide. Here, the PSA is given both permission and an obligation to monitor the environment in module X if a crewmember is not currently doing so. Similar policies could come into play when a crew member becomes overloaded or injured, such that he no longer is able to perform the task within predetermined criteria. In such cases, agents could be authorized and/or obligated to assist. For a more

complete discussion of adjustable autonomy, mixed-initiative interaction, and the role of policy, see [16.112, 16.120].

16.4.2 Cognitive and Robotic Prostheses

For some researchers, the ultimate in human-agent teamwork is the notion of agents that can function as extensions of the human brain cognitive prostheses and body robotic prostheses [16.39, 16.49, 16.52]. In this section we briefly suggest some preliminary considerations relating to human-agent interaction in the development of such capabilities.

At the outset, we recognize humans are an advantaged lot, each of us having been endowed with a "good brain and an unspecialized body" [16.68], which means that we are in a better position than any other creature to make and use a variety of tools. Moreover, bipedal locomotion has always had the beneficial side effect of freeing one hand to explore the environment and the other to wield those tools. Ford et al. argue that the accumulated tools of human history can all profitably be regarded as prostheses, not in the sense that they compensate for the specific disabilities of any given individual ([16.31]), but rather because they enable us to overcome the biological limitations shared by all of us: with reading and writing anyone can transcend the finite capacity of human memory; "with a power screwdriver anyone can drive the hardest screw; with a calculator, anyone can get the numbers right; with an aircraft anyone can fly to Paris; and with Deep Blue, anyone can beat the world chess champion" [16.39].

The prosthetic perspective can be contrasted with the traditional focus of Artificial Intelligence (AI) on standalone machine competence and its resulting preoccupation with the Turing Test as its measure of success [16.41]. Instead, argues Ford, we should start from a human-centered perspective. This implies that we must shift our goal "from making artificial superhumans who can replace us to making superhumanly intelligent artifacts that can amplify and support our own cognitive abilities" [16.49]. We don't need to jettison the acronym of AI, so long as we now take it to refer to the human's *Augmented* Intelligence.[30]

[30] The 1962 report of Engelbart entitled *Augmenting Human Intellect* presciently stressed the theme of "improving the intellectual effectiveness of the individual human being... through extensions of means developed and used in the past to help man apply his native sensory, mental, and motor capabilities. [Like] most systems its performance can best be improved by considering the whole as a set of interacting components rather than by considering the components in isolation" [16.36].

Eyeglasses, a well-known example of an ocular prosthesis,[31] provide a particularly useful example of three foundational concepts that are important to an understanding of cognitive and robotic prostheses:

1. *Transparency.* "Eyeglasses leverage and extend our ability to see, but in no way model our eyes: They don't look or act like them and wouldn't pass a Turing test for being an eye" [16.49]. A key feature of eyeglasses is that they can be used more or less transparently—by our forgetting they are present—just as humans with myopia don't think constantly about the wearing of the contact lenses but rather about the fact that they are seeing more effectively *through* them.[32]
2. *Unity.* Since our goal is not making smart eyeglasses but, rather, augmenting the human's ability to see, the minimum unit of discussion for the design of a prosthesis includes the device, the human being, and the environment in which the human will use the device. This mode of analysis necessarily blurs the line between humans and technology.[33]
3. *Fit.* Your eyeglasses won't fit me; neither will mine do you any good. Prostheses must fit the human and technological components together in ways that synergistically exploit their mutual strengths and mitigate their respective limitations. This implies a requirement for rich knowledge of how humans function.[34]

[31] The notion of augmenting sight through eyeglasses was "first mentioned by Roger Bacon in 1268. In the 1665 preface to Micrographia, Robert Hooke goes further, suggesting the addition 'of artificial Organs to the natural... to improve our other senses of hearing, smelling, tasting, and touching" [16.87].

[32] The manner in which perception operates during the use of good tools was insightfully described many years ago by Polanyi: "When we use a hammer to drive a nail, we attend to both nail and hammer, but in a different way. We watch the effect of our strokes on the nail and try to wield the hammer so as to hit the nail most effectively. When we bring down the hammer we do not feel that its handle has struck our palm but that its head has struck the nail. Yet in a sense we are certainly alert to the feelings in our palm and the fingers that hold the hammer. They guide us in handling it effectively, and the degree of attention that was given to the nail is given to the same extent but in a different way to these feelings. The difference may be stated by saying that the latter are not, like the nail, objects of our attention, but instruments of it. They are not watched in themselves; we watch something else while keeping intensely aware of them. I have a subsidiary awareness of the feeling in the palm of my hand which is merged into my focal awareness of my driving in the nail" [16.75].

[33] In 1960, Licklider [16.60] introduced the concept of *man-computer symbiosis:* "the hope is that, in not too many years, human brains and computing machines will be coupled together very tightly and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."

[34] A good example of this is the OZ cockpit display [16.49]. Through a groundbreaking study on the limits of human central and peripheral vision, IHMC's David Still discovered that peripheral vision can pick up 10 times the amount of detail than previously thought. Using this finding, he tailored the design of stimuli in a cockpit display to exploit the human sensory system's natural filter-

The elaboration of foundational concepts that are important to an understanding of cognitive and robotic prostheses, and the study of human functions in particular environments, happily dovetail with progress in the miniaturization of computing devices and the formulation of design principles for wearable computing [16.74]. Mann [16.67] was among the first to elucidate some of the necessary criteria for devices to be successfully subsumed into the human being's eudaemonic space (i.e., where the device seems to be part of the person).[35] He describes three required operational modes for wearable computing:

- **Constancy:** The computer runs continuously, and is *always ready* to interact with the user. Unlike a handheld device, laptop computer, or PDA, it does not need to be opened up and turned on prior to use. The signal flow from human to computer, and computer to human, ... runs continuously to provide a constant user interface.
- **Augmentation:** Traditional computing paradigms are based on the notion that computing is the primary task. Wearable computing, however, is based on the notion that computing is *not* the primary task. The assumption of wearable computing is that the user will be doing something else at the same time as doing the computing. Thus the computer should serve to augment the intellect or augment the senses ...
- **Mediation:** Unlike [traditional computers], the wearable computer can encapsulate us. It doesn't necessarily need to completely enclose us. There are two aspects to this encapsulation:
  i. **Solitude:** It can function as an information filter, and allow us to block out material we might not wish to experience, ... [or] it may simply allow us to alter our perception of reality.
  ii. **Privacy:** Mediation allows us to block or modify information leaving our encapsulated space. In the same way that ordinary clothing prevents others from seeing our naked bodies, the wearable computer may, for example, serve as an intermediary for interacting with untrusted systems.

Because of its ability to encapsulate us, ... [wearable computing devices] may also be able to make measurements of various physiological quantities.

Besides these operational modes, Mann describes six attributes of wearable systems:

ing and processing capabilities and to manipulate the data so it provides exactly what the pilot needs to know at any particular time. Stunningly, the OZ cockpit display is completely void of dials and gauges of ordinary cockpits, yet it is easier to learn, more straightforward to control, and more robust to temporary visual system impairment.

[35] Mann's formulations have evolved over time. Here, we discuss the version that can be found at http://www.eyetap.org/defs/glossary/wearcomp. See also Thad Starner's thorough survey of the field in his dissertation on Wearable Computing [16.87].

- **Unmonopolizing** of the user's attention, ... [though it may] mediate (augment, alter, or deliberately diminish) the sensory capabilities.
- **Unrestrictive** to the user: ... 'you can do other things while using it' ...
- **Observable** by the user: it can get your attention continuously if you want it to ...
- **Controllable** by the user ...
- **Attentive** to the environment: it is environmentally aware, multimodal, multi-sensory ... [thus increasing] the user's [situation] awareness ...
- **Communicative** to others: it can be used as a communications medium ...

DARPA's Augmented Cognition (AugCog) Program (http://www.darpa.mil/ito/research/ac/) is an example of an early effort focused on appropriately exploiting and integrating all available channels of communication from agents to humans (e.g., visual, auditory, tactile), and conversely sensing and interpreting a wide range of physiological measures of the human being in real-time so they can be used to tune agent behavior, and thus enhance joint human-machine performance.[36] For example, in IHMC's Adaptive Multi-Sensory Integration (AMI), augmented cognition prototype sets of system sensor agents (e.g., joystick), human sensor agents (e.g., EEG, pupil tracking, arousal meter), human display agents (e.g., visual, auditory, tactile), and adaptive automation agents (e.g., performing specific flight tasks) could work together with a pilot to promote stable and safe flight, sharing and adjusting aspects of control among the human and virtual crew member agents while taking system failures and human attention and stress loads into account [16.112].

While it is still too early to gauge the success of efforts such as AugCog, let alone to prescribe detailed principles for making cognitive and robotic prostheses acceptable to humans, it is clear that such modes of interaction will require new ways of thinking about human-agent interaction. In an insightful essay called *The Teddy* [16.71], Norman discusses some of the issues and implications of the widespread long-term habitual use of such technologies:

- Would we get so dependent that we would become disoriented without them?
- If they are constantly recording every event, should we allow them to be turned off? To protect civil liberties, you must be able to, and an indicator must show if someone's device is listening to you.

[36] A related program focused on similar issues with a robotics emphasis is NSF's Robotics and Human Augmentation (http://www.interact.nsf.gov/cise/descriptions.nsf/5b8c6c912ebf7f9b8525662c00723201/5e8661fa698fe674852565d9005985ef?OpenDocument). See also DARPA's Mobile Autonomous Robot Software (MARS) Robotic Vision 2020 Program (http://www.darpa.mil/ito/solicitations/FBO_02-15.html).

– Should it be programmed to always be supportive and encouraging (thus removing us from reality), or to give criticism and correction (thus resembling a nagging parent)? Getting the right balance is difficult in human relationships, how can we expect technology designers to do better?
– If we are never alone, when would we think? Would this accelerate the already tuned-out tendencies of headphone wearers?

16.5 Conclusions

In this chapter, we have outlined some of the technical and social challenges in the problem of making agents acceptable to people and have given examples and explanations of how a policy-based approach might be used to address some of those challenges. We hope that these initial efforts will inspire others to devote greater attention to reusable models and tools to assure the security, safety, naturalness, and effectiveness of human-agent teams.

Acknowledgements

The authors gratefully acknowledge the support of the DARPA CoABS, EPCA(CALO), Augmented Cognition, DAML, and Ultra*Log Programs, the NASA Cross-Enterprise and Intelligent Systems Programs, the Army Research Lab, the Office of Naval Research, the National Technology Alliance, and Fujitsu Labs while preparing this paper. We are also grateful for the contributions of James Allen, Alessandro Acquisti, Mike Bennett, Guy Boy, Kathleen Bradshaw, Mark Burstein, Murray Burke, Alberto Canas, Nate Chambers, Bill Clancey, Rob Cranfill, Cranfill, Grit Denker, Gary Edwards, Rich Feiertag, Ken Ford, Lucian Galescu, Yuri Gawdiak, Mike Goodrich, Mark Greaves, David Gunning, Jack Hansen, Pat Hayes, Mark Hoffman, Wayne Jansen, Hyuckchul Jung, Jim Just, Mike Kerstetter, Mike Kirton, Shri Kulkarni, Henry Lieberman, James Lott, Frank McCabe, Cindy Martin, Robin Murphy, Nicola Muscettola, Jerry Pratt, Debbie Prescott, Timothy Redmond, Sue Rho, Sebastien Rosset, Dylan Schmorrow, Debbie Schrekenghost, Kent Seamons, Mike Shafto, Maarten Sierhuis, Milind Tambe, Austin Tate, William Taysom, Ron Van Hoof, and Tim Wright.

References

16.1 A. Acquisti, M. Sierhuis, W. J. Clancey, J.M. Bradshaw: Agent-based modeling of collaboration and work practices onboard the International Space Station. *Proceedings of the Eleventh Conference on Computer-Generated Forces and Behavior Representation* (Orlando, FL 2002)

16.2 J. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, A. Stent: An architecture for a generic dialogue shell. Journal of Natural Language Engineering, 6(3), 1-16 (2000)
16.3 J.F. Allen, D.K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, A. Stent: Towards conversational human-computer interaction. AI Magazine, 22(4), 27-35 (2001)
16.4 D. Allsopp, P. Beautement, J.M. Bradshaw, E. Durfee, M. Kirton, C. Knoblock, N. Suri, A. Tate, C. Thompson: Coalition Agents eXperiement (CoAX): Multi-agent cooperation in an international coalition setting. In: A. Tate, J. Bradshaw, M. Pechoucek (eds.), *Special issue of IEEE Intelligent Systems* 17(3), 26-35 (2002)
16.5 R.Ambrose, C. Culbert, F. Rehnmark: An experimental investigation of dexterous robots using Eva tools and interfaces. *AIAA, 4593* (2001)
16.6 I. Asimov: Runaround. In: I. Asimov (ed.), *I, Robot.* pp. 33-51. London, England: Grafton Books. Originally published in *Astounding Science Fiction, 1942* pp. 94-103 (1942/1968)
16.7 K.S. Barber, M. Gamba, C. E. Martin: Representing and analyzing adaptive decision-making frameworks. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), *Agent Autonomy* (Dordrecht, The Netherlands, Kluwer, 2002) pp. 23-42
16.8 K.S. Barber, C.E. Martin: Agent autonomy: Specification, measurement, and dynamic adjustment. *Proceedings of the Workshop on Autonomy Control Software, International Conference on Autonomous Agents* (Seattle, WA, 1999)
16.9 M. Boman: Norms in artificial decision-making. Artificial Intelligence and Law, 7, 17-35 (1999)
16.10 G. Boy: Human-centered design of artificial agents: A cognitive function analysis approach. In: J.M. Bradshaw (ed.), *Handbook of Software Agents.* (Cambridge, MA, AAAI Press/The MIT Press, 2002) (in press)
16.11 J.M. Bradshaw, G. Boy, E. Durfee, M. Gruninger, H. Hexmoor, N. Suri, M. Tambe, M. Uschold, J. Vitek (eds.): *Software Agents for the Warfighter. ITAC Consortium Report* (Cambridge, MA, AAAI Press/The MIT Press 2002)
16.12 J.M. Bradshaw, S. Dutfield, P. Benoit, J.D. Woolley: KAoS: Toward an industrial-strength generic agent architecture. In: J.M. Bradshaw (ed.), *Software Agents* (Cambridge, MA, AAAI Press/The MIT Press 1997) pp. 375-418
16.13 J.M. Bradshaw, M. Greaves, H. Holmback, W. Jansen, T. Karygiannis, B. Silverman, N. Suri, A. Wong: Agents for the masses: Is it possible to make development of sophisticated agents simple enough to be practical? IEEE Intelligent Systems (March-April), 53-63 (1999)
16.14 J.M. Bradshaw, M. Sierhuis, A. Acquisti, P. Feltovich, R. Hoffman, R. Jeffers, D. Prescott, N. Suri, A. Uszok, R. Van Hoof: Adjustable autonomy and human-agent teamwork in practice: An interim report on space applications. In: H. Hexmoor, R. Falcone, C. Castelfranchi (eds.), *Agent Autonomy* pp. 243-280
16.15 J.M. Bradshaw, N. Suri, M.R. Breedy, A. Canas, R. Davis, K.M. Ford, R. Hoffman, R. Jeffers, S. Kulkarni, J. Lott, T. Reichherzer, A. Uszok: Terraforming cyberspace. In: D.C. Marinescu, C. Lee (eds.), *Process Coordination and Ubiquitous Computing* pp. 165-185. Boca Raton, FL: CRC Press. Updated and expanded version of an article that originally appeared in IEEE Intelligent Systems, July 2001, pp. 49-56
16.16 J.M. Bradshaw, A. Uszok, R. Jeffers, N. Suri, P. Hayes, M.H. Burstein, A. Acquisti, B. Benyo, M.R. Breedy, M. Carvalho, D.Diller, M. Johnson, S. Kulkarni, J. Lott, M. Sierhuis, R. Van Hoof: Representation and reasoning for DAML-based policy and domain services in KAoS and Nomads. *Proceedings of the Autonomous Agents and Multi-Agent Systems Conference (AAMAS 2003)* (Melbourne, Australia, New York, NY: ACM Press)

16.17 S.Brainov, H. Hexmoor: Quantifying autonomy. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), *Agent Autonomy.* (Dordrecht, The Netherlands, Kluwer) pp. 43-56
16.18 C. Breazeal, B. Scassellati: How to build robots that make friends and influence people. *IROS* (Kyonjiu, Korea)
16.19 R.R. Burridge, J. Graham, K. Shillcutt, R. Hirsh, D. Kortenkamp: Experiments with an EVA assistant robot. *Proceedings of the Seventh International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)* (Nara, Japan)
16.20 M.H. Burstein, D. V. McDermott: Issues in the development of human-computer mixed-initiative planning. In: B. Gorayska J.L. Mey (eds.), *Cognitive Technology: In Search of a Humane Interface.* (Elsevier Science)
16.21 K. Capek: R. U. R. (Rossum's Universal Robots). In: P. Kussi (ed.), *Toward the Radical Center: A Karel Capek Reader* (North Haven, CT: Catbird Press)
16.22 H. Chalupsky, Y. Gil, C. A. Knoblock, K. Lerman, J. Oh, D.V. Pynadath, T.A. Russ, M. Tambe: Electric Elves: Agent technology for supporting human organizations. AI Magazine, 2, pp. 11-24
16.23 W.J. Clancey: Simulating activities: Relating motives, deliberation, and attentive coordination. *Cognitive Systems Review, special issue on Situated and Embodied Cognition*
16.24 H.H. Clark: *Arenas of Language Use* (Chicago, IL: University of Chicago Press)
16.25 R. Clarke: Asimov's laws of robotics: Implications for information technology, Parts 1 and 2. IEEE Computer, pp. 53-66.
16.26 J. Clute, P. Nicholls: Grolier Science Fiction: The Multimedia Encyclopedia of Science Fiction (CD-ROM). In: Danbury, CT: Grolier Electronic Publishing
16.27 P. Cohen, R. Coulston, K. Krout: Multimodal interaction during multiparty dialogues: Initial results. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces* (Pittsburgh, PA)
16.28 P.R. Cohen, H.J. Levesque: *Teamwork.* Technote 504. Menlo Park, CA, SRI International, March
16.29 R. Cohen, C. Allaby, C. Cumbaa, M. Fitzgerald, K. Ho, B. Hui, C. Latulipe, F. Lu, N. Moussa, D. Pooley, A. Qian, S. Siddiqi: What is initiative? *User Modeling and User-Adapted Interaction* 8(3-4), 171-214
16.30 R. Cohen, M. Fleming: Adjusting the autonomy in mixed-initiative systems by reasoning about interaction. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), *Agent Autonomy* (Dordrecht, The Netherlands: Kluwer) pp. 105-122
16.31 E. Cole, P. Dehdashti: Computer-based cognitive prosthetics: Assistive technology for the treatment of cognitive disabilities. *Proceedings of the Third International ACM Conference on Assistive Technologies (ACM SIGCAPH - Computers and the Physically Handicapped)* (Marina del Rey, CA)
16.32 R. Conte, C. Castelfranchi: *Cognitive and social action* (London, England: UCL Press)
16.33 M. d'Inverno, M. Luck: *Understanding Agent Systems* (Berlin, Germany, Springer-Verlag)
16.34 N. Damianou, N. Dulay, E.C. Lupu, M.S. Sloman: *Ponder: A Language for Specifying Security and Management Policies for Distributed Systems, Version 2.3.* Imperial College of Science, Technology and Medicine, Department of Computing, 20 October 2000
16.35 G. Dorais, R.P. Bonasso, D. Kortenkamp, B. Pell, D. Schrekenghost: Adjustable autonomy for human-centered autonomous systems on Mars. *Proceedings of the AAAI Spring Symposium on Agents with Adjustable Autonomy.*

AAAI Technical Report SS-99-06 (Menlo Park, CA, Menlo Park, CA, AAAI Press)
16.36 D. C. Engelbart: *Augmenting Human Intellect: A Conceptual Framework.* Air Force Office of Scientific Research, AFOSR-3233 Summary Report, SRI Project 3578 (Stanford Research Institute, October)
16.37 R. Falcone, C. Castelfranchi: From automaticity to autonomy: The frontier of artificial agents. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), *Agent Autonomy.* (Dordrecht, The Netherlands, Kluwer) pp. 79-103
16.38 G. Ferguson, J. Allen, B. Miller: TRAINS-95: Towards a mixed-initiative planning assistant. *Proceedings of the Third Conference on Artificial Intelligence Planning Systems (AIPS-96)* (Edinburgh, Scotland) pp. 70-77
16.39 K.M. Ford, C. Glymour, P. Hayes: Cognitive prostheses. AI Magazine, 18(3), 104
16.40 K.M. Ford, C. Glymour, P.J. Hayes (eds.): *Android Epistemology* (Menlo Park, CA: AAAI Press / The MIT Press)
16.41 K.M. Ford, P. Hayes: On computational wings: Rethinking the goals of Artificial Intelligence. *Scientific American. Special issue on Exploring Intelligence* 9 (4), 78-83
16.42 I. Foster, C. Kesselman (eds.): *The Grid: Blueprint for a New Computing Infrastructure.* (San Francisco, CA, Morgan Kaufmann)
16.43 J. Fox, S. Das: *Safe and Sound: Artificial Intelligence in Hazardous Applications* (Menlo Park, CA: AAAI Press/The MIT Press)
16.44 Y. Gawdiak, J.M. Bradshaw, B. Williams, H. Thomas: R2D2 in a softball: The Personal Satellite Assistant. In: H. Lieberman (ed.), *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI 2000)*(New Orleans, LA, New York: ACM Press) pp. 125-128
16.45 N.A. Gershenfeld: *When Things Start to Think* (New York, Henry Holt and Company)
16.46 M.A.Goodrich, D.R. Olsen Jr., J.W. Crandall, T.J. Palmer: Experiments in adjustable autonomy. *Proceedings of the IJCAI_01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents.* Seattle, WA, Menlo Park, CA, AAAI Press
16.47 M. Greaves, H. Holmback, J.M. Bradshaw: What is a conversation policy? M. Greaves, J.M. Bradshaw (eds.), *Proceedings of the Autonomous Agents '99 Workshop on Specifying and Implementing Conversation Policies* (Seattle, WA) pp. 1-9
16.48 M. Greaves, H. Holmback, J.M. Bradshaw: Agent conversation policies. In: J.M. Bradshaw (ed.), *Handbook of Agent Technology.* (pp. In: preparation) (Cambridge, MA, AAAI Press/The MIT Press)
16.49 S. Hamilton: Thinking outside the box at IHMC. IEEE Computer, 61-71
16.50 P.A. Hancock, S.F. Scallen: Allocating functions in human-machine systems. In: R. Hoffman, M.F. Sherrick, J.S. Warm (eds.), *Viewing Psychology as a Whole* (Washington, D.C., American Psychological Association) pp. 509-540
16.51 R. Hoffman, P. Feltovich, K.M. Ford, D.D. Woods, G. Klein, A. Feltovich: A rose by any other name? would probably be given an acronym. IEEE Intelligent Systems, July-August, pp. 72-80
16.52 R. R. Hoffman, K.M. Ford, P.J. Hayes, J.M. Bradshaw: The Borg hypothesis. IEEE Intelligent Systems (in press)
16.53 E. Horvitz: Principles of mixed-initiative user interfaces. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'99)* (Pittsburgh, PA, New York: ACM Press)

16.54 E. Horvitz, A. Jacobs, D. Hovel: Attention-sensitive alerting. *Proceedings of the Conference on Uncertainty and Artificial Intelligence (UAI'99)* (Stockholm, Sweden) pp. 305-313
16.55 B. Joy: Why the future doesn't need us. *Wired*, 4
16.56 A. Kay: User interface: A personal view. In: B. Laurel (ed.), *The Art of Human-Computer Interface Design* (Reading, MA, Addison-Wesley) pp. 191-208
16.57 G. Knoll, N. Suri, J.M. Bradshaw: Path-based security for mobile agents. *Proceedings of the First International Workshop onthe Security of Mobile Multi-Agent Systems (SEMAS-2001) at the Fifth International Conference on Autonomous Agents (Agents 2001)* (Montreal, CA, New York: ACM Press)pp. 54-60
16.58 A. Kolber: *Defining Business Rules: What Are They Really?* Revision 1.3, Final Report. Business Rules Group (formerly the GUIDE Business Rules Project, July)
16.59 N.G. Leveson: *Safeware: System Safety and Computers: A Guide to Preventing Accidents and Losses Caused by Technology* (Boston, MA, Addison-Wesley)
16.60 J.C.R. Licklider: Man-computer symbiosis. *IRE Transactions in Electronics. New York: Institute of Radio Engineers.*, 4-11
16.61 H. Lieberman (ed.): *Your Wish is My Command: Programming By Example* (San Francisco, CA: Morgan Kaufmann)
16.62 F. Lopez y Lopez, M. Luck, M. d'Inverno: A framework for norm-based inter-agent dependence. *Proceedings of the Third Mexican Internation Conference on Computer Science*
16.63 F. Lopez, Y. Lopez, M. Luck, M. d'Inverno: Constraining autonomy through norms. *Proceedings of the Conference on Autonomous Agents and Multi-Agent Systems* (Bologna, Italy) pp. 674-681
16.64 S. Lubar: *InfoCulture: The Smithsonian Book of Information and Inventions* (Boston, MA, Houghton Mifflin Company)
16.65 M. Luck, M. D'Inverno, S. Munroe: Autonomy: Variable and generative. In: H. Hexmoor, C. Castelfranchi, R. Falcone (ed.), *Agent Autonomy* (Dordrecht, The Netherlands, Kluwer) pp. 9-22
16.66 P. Maes: Agents that reduce work and information overload. In: J.M. Bradshaw (ed.), *Software Agents* (Cambridge, MA, AAAI Press/The MIT Press 1997) pp. 145-164
16.67 S. Mann: Wearable computing: A first step toward personal imaging. IEEE Computer, 30(2), 25-32 (1997)
16.68 D. Morris: *Peoplewatching* (London, England, Vintage 2002)
16.69 R.R. Murphy: *Introduction to AI Robotics* (Cambridge, MA, The MIT Press 2000)
16.70 D.A. Norman: The 'problem' with automation: Inappropriate feedback and interaction, not over-automation. In: D.E. Broadbend, J. Reason, A. Baddeley (eds.), *Human Factors in Hazardous Situations* (Oxford, England: Clarendon Press 1990) pp. 137-145
16.71 D.A. Norman: *Turn Signals are the Facial Expressions of Automobiles* (Reading, MA, Addison-Wesley 1992)
16.72 D.A. Norman: How might people interact with agents? In: J.M. Bradshaw (ed.), *Software Agents.* (Cambridge, MA, The AAAI Press/The MIT Press 1997) pp. 49-55
16.73 S.L. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, D. Ferro: Designing the user

interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. Human Computer Interaction, 15(4), 263-322 (2000)
16.74 A.P. Pentland: Wearable intelligence. *Scientific American. Special issue on Exploring Intelligence*, 9(4), pp. 90-95 (1998)
16.75 M. Polanyi: *Personal Knowledge: Toward a Post-Critical Philosophy* (Chicago, IL, The University of Chicago Press 1962)
16.76 D. Pynadath, M. Tambe: Revisiting Asimov's first law: A response to the call to arms. *Proceedings of ATAL 01* (2001)
16.77 N. Sarter, D.D. Woods, C.E. Billings: Automation surprises. In: G. Salvendy (ed.), *Handbook of Human factors/Ergonomics, 2nd Edition* (New York, NY: John Wiley 1997)
16.78 R.K. Sawyer: *Creating Conversations: Improvisation in Everyday Discourse* (Cresskill, NJ, Hampton Press 2001)
16.79 P. Scerri, D. Pynadath, M. Tambe: Adjustable autonomy for the real world. In: R. Falcone (ed.), *Agent Autonomy* (Dordrecht, The Netherlands: Kluwer 2002) pp. 163-190
16.80 P. Schelde: *Androids, Humanoids, and Other Science Fiction Monsters* (New York, New York University Press 1993)
16.81 D. Schreckenghost, C. Martin, P. Bonasso, D. Kortenkamp, T. Milam, C. Thronesbery: Supporting group interaction among humans and autonomous agents. *Submitted for publication* (2003)
16.82 D. Schreckenghost, C. Martin, C. Thronesbery: Specifying organizational policies and individual preferences for human-software interaction. *Submitted for publication* (2003)
16.83 K.E. Seamons, M. Winslet, T. Yu: Limiting the disclosure of access control policies during automated trust negotiation. *Proceedings of the Network and Distributed Systems Symposium* (2001)
16.84 Y. Shoham, M. Tennenholtz: On the synthesis of useful social laws for artificial agent societies. *Proceedings of the Tenth National Conference on Artificial Intelligence* (San Jose, CA 1992) pp. 276-281
16.85 M. Sierhuis, J.M. Bradshaw, A. Acquisti, R. Van Hoof, R. Jeffers, A. Uszok: Human-agent teamwork and adjustable autonomy in practice. *Proceedings of the Seventh International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)* (Nara, Japan 2003)
16.86 W.J. Smith: *The Behavior of Communicating* (Cambridge, MA, Harvard University Press 1977)
16.87 T.E. Starner: *Wearable Computing and Contextual Awareness*. Doctor of Philosophy, Massachusetts Institute of Technology (1999)
16.88 N. Suri, J.M. Bradshaw, M.R. Breedy, P.T. Groth, G.A. Hill, R. Jeffers: Strong Mobility and Fine-Grained Resource Control in NOMADS. *Proceedings of the 2nd International Symposium on Agents Systems and Applications and the 4th International Symposium on Mobile Agents (ASA/MA 2000)* (Zurich, Switzerland, Berlin: Springer-Verlag 2000)
16.89 N. Suri, J.M. Bradshaw, M.R. Breedy, P.T. Groth, G.A. Hill, R. Jeffers, T.R. Mitrovich, B.R. Pouliot, D.S. Smith: NOMADS: Toward an environment for strong and safe agent mobility. *Proceedings of Autonomous Agents 2000* (Barcelona, Spain, New York: ACM Press 2000)
16.90 N. Suri, J.M. Bradshaw, M.H. Burstein, A. Uszok, B. Benyo, M.R. Breedy, M. Carvalho, D. Diller, P.T. Groth, R. Jeffers, M. Johnson, S. Kulkarni, J. Lott: OWL-based policy enforcement for semantic data transformation and filtering in multi-agent systems. *Proceedings of the Autonomous Agents and*

Multi-Agent Systems Conference (AAMAS 2003) (Melbourne, Australia, New York, NY, ACM Press 2003)

16.91 N. Suri, J.M. Bradshaw, M. Carvalho, M.R. Breedy, T.B. Cowin, R. Saavendra, S. Kulkarni: Applying agile computing to support efficient and policy-controlled sensor information feeds in the Army Future Combat Systems environment. *Proceedings of the Annual U.S. Army Collaborative Technology Alliance (CTA) Symposium* (2003)

16.92 N. Suri, J.M. Bradshaw, M. Carvalho, T.B. Cowin, M.R. Breedy, P. T. Groth, R. Saavendra: Agile computing: Bridging the gap between grid computing and ad-hoc peer-to-peer resource sharing. In: O.F. Rana (ed.), *Proceedings of the Third International Workshop on Agent-Based Cluster and Grid Computing* (Tokyo, Japan 2003)

16.93 N. Suri, J.M. Bradshaw, M.R. Breedy, P.T. Groth, G.A. Hill, R. Jeffers, T.R. Mitrovich, B.R. Pouliot, D.S Smith: NOMADS: Toward an environment for strong and safe agent mobility. *Proceedings of Autonomous Agents 2000* (Barcelona, Spain, ACM Press, NY 2000)

16.94 N. Suri, M. Carvalho, J.M. Bradshaw, M.R. Breedy, T.B. Cowin, P.T. Groth, R. Saavendra, A. Uszok: Mobile code for policy enforcement. *Policy 2003* (Como, Italy 2003)

16.95 M. Tambe, D. Pynadath, C. Chauvat, A. Das, G. Kaminka: Adaptive agent architectures for heterogeneous team members. *Proceedings of the International Conference on Multi-Agent Systems (ICMAS 2000)*

16.96 M. Tambe, W. Shen, M. Mataric, D.V. Pynadath, D. Goldberg, P.J. Modi, Z. Qiu, B. Salemi: Teamwork in cyberspace: Using TEAMCORE to make agents team-ready. *Proceedings of the AAAI Spring Symposium on Agents in Cyberspace* (Menlo Park, CA, Menlo Park, CA, The AAAI Press 1999)

16.97 C. Trevarthen: Communication and cooperation in early infancy: A description of primary intersubjectivity. In: M. Bullowa (ed.), *Before Speech* (Cambridge, England: Cambridge University Press 1979) pp. 321-348

16.98 A. Uszok, J.M. Bradshaw, P. Hayes, R. Jeffers, M. Johnson, S. Kulkarni, M.R. Breedy, J. Lott, L. Bunch: DAML reality check: A case study of KAoS domain and policy services. *Submitted to the Second International Semantic Web Conference (ISWC 03)* (Sanibel Island, Florida, October 2003)

16.99 A. Vanderbilt: *Amy Vanderbilt's New Complete Book of Etiquette: The Guide to Gracious Living* (Garden City, NY, Doubleday and Company 1952/1963)

16.100 H. Verhagen: Norms and artificial agents. *Sixth Meeting of the Special Interest Group on Agent-Based Social Simulation, ESPRIT Network of Excellence on Agent-Based Computing* (Amsterdam, Holland) http://abss.cfpm.org/amsterdam-01/abssnorms.pdf (2001)

16.101 D. Weld, O. Etzioni: The firsts law of robotics: A call to arms. *Proceedings of the National Conference on Artificial Intelligence (AAAI 94)* pp. 1042-1047 (1994)

16.102 M. Winslett, T. Yu, K.E. Seamons, A. Hess, J. Jacobson, R. Jarvis , B. Smith, L. Yu: Negotiating trust on the Web. IEEE Internet Computing, 30-37 (2002)

16.103 K.S. Barber, M. Gamba, C.E. Martin: Representing and analyzing adaptive decision-making frameworks. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), Agent Autonomy. (Dordrecht, The Netherlands, Kluwer) pp. 23-42 (2002)

16.104 G. Boella: Obligations and cooperation: Two sides of social rationality. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), Agent Autonomy. (Dordrecht, The Netherlands, Kluwer) pp. 57-78 (2002)

16.105 R. Cohen, M. Fleming: Adjusting the autonomy in mixed-initiative systems by reasoning about interaction. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), Agent Autonomy. (Dordrecht, The Netherlands, Kluwer) pp. 105-122 (2002)
16.106 R. Falcone, C. Castelfranchi: From automaticity to autonomy: The frontier of artificial agents. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), Agent Autonomy. (Dordrecht, The Netherlands, Kluwer) pp. 79-103 (2002)
16.107 P.A. Hancock, S.F. Scallen: Allocating functions in human-machine systems. In: R. Hoffman, M.F. Sherrick, J.S. Warm (eds.), Viewing Psychology as a Whole. (Washington, D.C., American Psychological Association) pp. 509-540 (1998)
16.108 E. Horvitz: Principles of mixed-initiative user interfaces. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'99)* (Pittsburgh, PA, New York, ACM Press 1999)
16.109 E. Horvitz, A. Jacobs, D. Hovel: Attention-sensitive alerting. *Proceedings of the Conference on Uncertainty and Artificial Intelligence (UAI'99)* (Stockholm, Sweden) pp. 305-313 (1999)
16.110 K. Myers, D. Morley: Directing agents. In: H. Hexmoor, C. Castelfranchi, R. Falcone (eds.), Agent Autonomy (Dordrecht, The Netherlands: Kluwer) pp. 143-162 (2003)
16.111 P. Scerri, D. Pynadath, M. Tambe: Adjustable autonomy for the real world. In: R. Falcone (ed.), Agent Autonomy (Dordrecht, The Netherlands: Kluwer) pp. 163-190 (2002)
16.112 J.M. Bradshaw, P. Feltovich, H. Jung, S. Kulkarni, W. Taysom, A. Uszok, (eds.): Dimensions of adjustable autonomy and mixed-initiative interaction. In: M. Klusch, G. Weiss, M. Rovatsos (eds.), Computational Autonomy (Berlin, Springer-Verlag 2004) (in press)
16.113 A. Uszok, J.M. Bradshaw, R. Jeffers, M. Johnson, A. Tate, J. Dalton, S. Aitken: Policy and contract management for semantic web services. *Proceedings of the AAAI Spring Symposium*, Standford, CA, AAAI Press, March 22-24 (in press)
16.114 A. Uszok, J.M. Bradshaw, R. Jeffers: KAoS: A policy and domain services framework for grid computing and grid computing and semantic web services. In: C. Jensen, S. Prslad, T. Dimitrakos (eds.), Trust Management: Second International Conference (iTrust 2004) Proceedings, Oxford, UK, March/April, Lecture Notes in Computer Science 2995, Berlin, Springer, pp. 16-26
16.115 P. Feltovich, J.M. Bradshaw, R. Jeffers, N. Suri, A. Uszok: Social order and adaptability in animal and human cultures as analogues for agent communities: Toward a policy-based approach. In: A. Omicin, P. Petta, J. Pitt (eds.), Engineering Societies in the Agents World IV. Lecture Notes in Computer Science Series. Berlin, Germany, Springer-Verlag (in press)
16.116 J. Barwise, J. Perry: Situations and Attitudes. Cambridge, MA: MIT Press (1983)
16.117 K. Devlin: Logic and Information. Cambridge, England: Cambridge University Press (1991)
16.118 J.J. Gibson: The Ecological Approach to Visual Perception. Boston, MA: Houghton Mifflin (1979)
16.119 D.A. Norman: Cognitive artifacts. In: J.M. Carroll (ed.), Designing Interaction: Psychology at the Human-Computer Interface. (pp. 17-38). Cambridge: Cambridge University Press (1992)
16.120 J.M. Bradshaw, H. Jung, S. Kulkarni, J. Allen, L. Bunch, N. Chambers, P. Feltovich, L. Galescu, R. Jeffers, M. Johnson, R. Taysom, A. Uszok: Toward

trustworthy adjustable autonomy and mixed-initiative interaction in KAoS. Submitted for publication (2004)
16.121 G. Tonti, J.M. Bradshaw, R. Jeffers, R. Montanari, N. Suri, A. Uszok: Semantic Web languages for policy representation and reasoning: A comparison of KAoS, Rei, and Ponder. In: D. Fensel, K. Sycara, J. Mylopoulos (eds.), The Semantic WebISWC 2003. Proceedings of the Second International Semantic Web Conference, Sanibel Island, Florida, USA, October 2003, LNCS 2870. (pp. 419-437). Berlin: Springer (2003)
16.122 J. Lott, J.M. Bradshaw, A. Uszok, R. Jeffers: KAoS policy management for control of security mechanisms in a large-scale distributed system (submitted for publication, 2004)