Social Order and Adaptability in Animal and Human Cultures as Analogues for Agent Communities: Toward a Policy-Based Approach

Paul J. Feltovich, Jeffrey M. Bradshaw, Renia Jeffers, Niranjan Suri, Andrzej Uszok
Institute for Human and Machine Cognition/University of West Florida
40 S. Alcaniz, Pensacola, FL 32501
{pfeltovich, jbradshaw, rjeffers, nsuri, auszok}@ihmc.us

Abstract. In this paper we discuss some of the ways social order is maintained in animal and human realms, with the goal of enriching our thinking about mechanisms that might be employed in developing similar means of ordering communities of agents. We present examples from our current work in human-agent teamwork, and we speculate about some new directions this kind of research might take. Since communities also need to change over time to cope with changing circumstances, we also speculate on means that regulatory bodies can use to adapt.

1. Introduction

As computational systems with increasing autonomy interact with humans in more complex ways—and with the welfare of the humans sometimes dependent on the conduct of the agents—there is a natural concern that the agents act in ways that are acceptable to people [7; 51]. In addition to traditional concerns for safety and robustness in such systems [12], there are important social aspects relating to predictability, control, feedback, order, and naturalness of the interaction that must be attended to [8; 10; 50]. In this paper we investigate just some of the ways social order is maintained in animal and human realms (sections 2 and 3), with the goal of enriching our thinking about mechanisms that might be employed to enhance order in mixed human-agent teams.[1] We present examples of such systems that have been created to support agent-based applications (section 4), and we speculate about new directions this kind of research might take (section 5). Since enduring communities also need to change over time to cope with changing circumstances, we speculate briefly on means that regulatory bodies can utilize for supporting adaptation (section 6). Finally, we present some concluding observations (section 7).

[1] In this sense, we agree with the conjecture of Norman: “Technology recapitulates phylogeny” [50, p. 134].

2. Some Sources of Order in the Animal World

We start by examining some of the ways that animals cooperate and maintain order. Why would individuals ever choose to cooperate with others to pursue their aims, rather than "going it alone"? In the animal realm, ethnologists and evolutionary biologists have taken a fairly common stance with regard to this question. Speaking of the process of mutual "attunement" (roughly, "getting to know one another") among individuals, a component process of cooperation, biologist W.J. Smith states:

> *Such attunement is necessary when no single individual can fully control an encounter—when participants in encounters must depend on each other for a useful outcome. The value of that outcome need not be equal for each participant, but it must exceed for each the average payoff that would come from eschewing the interaction [61, p. 366].*

Smith goes on to discuss two main benefits that accrue from such processes of cooperation or "joint activity." The first is that certain tasks get accomplished that could not have been accomplished by any individual. The second is that these kinds of activities, over time, yield increased *inter-predictability* among the parties; they come to know each other's ways. This can have constructive benefits: for instance, knowledge of the other's capabilities might be tapped during future cooperation. It can also yield protective benefits: for example, learning the other's "hot buttons" that tend to invoke hostility. But the main benefit of predictability is the social order it contributes to the group. Gross, mutual unpredictability is almost definitional of *disorder*. Predictability and order are so important to animals that they seem to go to great lengths to build but also maintain it: For instance:

> *[Some male birds] remember how to recognize previous neighbors by their individually distinctive songs and remember the location in which each neighbor belongs. Relationships with known neighbors are valuable and those with strangers are problematic. Known mutual boundaries can be reestablished with much less effort and uncertainty than goes into the task of working out relationships with new neighbors [61, p. 365].*

Animals engage in joint activities, in which they get to know each other, in part through processes of signaling and display that are associated with predictable kinds of behaviors. That is, display and signaling behavior among animals supports joint activity by providing more or less rough clues to others concerning what each individual is about to do. Displays and signals can range widely in form (e.g., vocalizations, body posture, facial expressions):

> *Each individual has a repertoire of behavior made up of all the many kinds of acts it can perform. It can be thought of as continuously choosing among these acts, even at times when its behavior is unchanging (among the choices available at any instant is to do whatever was done in the previous instant). Any choice can be called a 'behavioral selection.'*
>
> *Each kind of display has a consistent and specifiable relationship to certain choices. It is performed in correlation with some kinds of behavior and not*

> *others. Thus, to know that an individual is performing a particular display is to learn something about the behavior it may select—every display can thus be described as encoding messages about behavioral selections [60, p. 87].*

Hence, display behavior has an anticipatory, predictive (but only a *probabilistically* predictive) function. It is a clue, sometimes highly indicative, sometimes much less so,[2] to what an individual is about to do. It also decouples actual action from a kind of notice that it is about to happen.[3] This decoupling both invites and enables others to participate in coordination, support, or avoidance with respect to what might occur. This joint engagement in an activity would not be possible if the activity were merely executed and not signaled in advance. In this sense, display is an important ingredient in enabling things like coordination and teamwork.

While signaling and display can take many and complicated forms, even in the animal world, biologist Smith has advanced ten signal-behavior couplings that appear to be pervasive in almost all vertebrates, although they might manifest different physical forms in different species [60, pp. 87-126]. The fact that these are so pervasive suggests they may be particularly fundamental. We will briefly describe each of these types of displays and signals along with possible functions they could serve within agent communities.

2.1. Interactional Displays

Interactional displays indicate availability or unavailability to participate in joint activity. These displays "primarily provide information about the communicator's readiness or lack of readiness, to join in acts that involve other individuals" [60, p. 88]. Since they may be associated with more than one kind of interaction, they do not specify any one kind. They might indicate readiness to copulate, associate, attack an intruder, and so forth. Hence, they are anticipatory to various kinds of intended joint activity, simply signaling a readiness (or lack thereof) to join in association with others.

This category also includes displays indicating *absence of opportunity to interact*. These displays essentially signal that an individual is alone and has nobody else to interact with, for example, when an individual is the last remaining at the nest or territory. This category also includes signals of *shunning interaction*. These are simply signals that the initiator does not want interaction with others, and this intention can range from mild to fierce.

Example interactional forms. Kinds of chirping. Various forms of bowing. "Tidbitting"—offering a morsel of food. Forms of touching. Signals from a

[2] Sometimes the ambiguity of the signal itself serves an important function, for example as an indicator that the signaler's next move may depend on the response its current move evokes.

[3] To see why this may be useful, consider the signaling functions of the lights on the back of a car: "[W]e use turn signals and brake lights to tell others of our actions and intentions. In the case of brake lights, we signal actions as we carry them out. In the case of turn signals, we signal our intentions before we actually commit them into action. In either case, we allow others to know our future actions so that we can ensure that there is no conflict" [50, p. 129].

subordinate to a dominant, the purpose of which is to test the dominant's willingness to interact, to tolerate interaction.
Absence of opportunity: Loud sounds, loud singing, howling (e.g., one jackal howls, and all the rest in the area howl in response), assuming high, visible physical positions, special kinds of flight patterns or displays.
Shunning: Interestingly, various forms of displaying the tongue. Chittering barks. Vocalizations at special, unusual frequencies.
Possible functions in agent communities. Displays in this general category clearly have benefits for coordination among groups of agents by providing information about which are or are not in a position to interact with others, in what ways, when, and so forth, e.g.: Call me. I am open for calls. I need to talk to someone. May I interject, may I say something?
Absence of opportunity: I am out of touch. I am working all alone. I have no help. I have lost contact with everybody.
Shunning: Do *not* attempt to communicate with me for whatever reason, e.g., my line is bugged, or I am involved in something that cannot be interrupted. Leave me alone.

While the general interactional displays just discussed are non-specific in the activity they portend, others are more specific.

2.2. Seeking Displays

Displays indicating that one is *seeking* joint activity are similar to the interactional ones in that they indicate a readiness to participate in some kind joint activity but differ in that they indicate *active attempt* at engaging in a particular kind of activity rather than just a general state of availability or receptiveness:

> *"Animals may display while seeking the opportunity to perform some kind of activity during what ethnologists call 'appetitive' behavior as distinguished from 'consummatory' behavior in which activity is completed. The behavioral selection about which a display provides information if it is done* only *in this way can be termed 'seeking.' What a communicator is seeking to do is encoded in the same display by a second behavioral selection message. The display is interpreted as providing not just the information that a communicator is ready to do this second selection, but that its behavior includes seeking or preparing to seek an opportunity" [60, p. 118].*

The seeking display can be associated with many kinds of activities, seeking, for example, to interact, associate, copulate, attack, or escape.
Example forms. These are associated with so many kinds of behaviors that their particular forms vary widely.
Possible functions in agent communities. Agents that indicate to others what they are trying to do can elicit the right form of aid from others, can contribute to possible coordination among tasks, and the like.

2.3. Receptiveness Displays

Displays indicating *receptiveness* are the inverse of seeking displays, i.e., they indicate a specific response to the seeking of particular kinds of activities by others:

> *"Some displays indicate the behavioral selections that a communicator will* accept*, not those it is prepared to perform. At least two behavioral selection messages must be provided by such a display, one indicating that the communicator will behave receptively and another indicating the class of acts to which it is receptive. Effectively, the communicator adopts the role of soliciting acts from another individual; it does not offer them"* [60, p. 122].

The display indicating receptiveness indicates that the communicator is willing to engage in a behavior, or set of behaviors, initiated by another. An interesting form of soliciting has to do with being receptive to "aid or care" and is common among infants who indicate receptiveness to feeding, grooming, shading, and so forth. Although often associated with the young, these displays sometimes carry over into adult relationships, as when a female mate solicits various forms of "help with the nest" from her male partner [60, p. 125].

Example forms. As with seeking displays, receptivity displays are so diverse that they defy general description.

Possible functions in agent communities. As with the seeking displays, receptivity displays can contribute to cooperation in the conduct of activity and to the coordination among activities.

2.4. Attack and Escape Displays

Displays indicating *attack* and *escape*:

> *"are said to encode either, or both, of attack and escape messages when all their occurrence is correlated with a range of attack- or escape-related behavior. Behavioral indices of attack differ among and within species, but include acts that, if completed, will harm another individual. Escape behavior can be any appropriate form of avoidance, ranging from headlong fleeing to turning aside, or even freezing and other ways of hiding" [60, p. 93].*

Attack and escape displays may differ, but they are sometimes more or less the same display, differing only in degree or subtle nuance. They have value both between and within groups, for instance, to muster help against an intruder or to avoid inadvertent flare-ups (e.g., one group member coming upon another by surprise). Various choreographies of interactive displays relating to attacking and escaping can more often than not serve to avoid actual combat. Actual fighting is more likely to happen among relatively unfamiliar groups [60, p. 94], partly because they have less mutual predictability, including prediction of each other's reaction to display activities that can fend off real fighting.

Example forms. Body posture and orientation. Head bobbing. Forms of jumping. Baring teeth.

Possible functions in agent communities. Displays in this category have increased importance when agents are acting in adversarial environments, such as those found in military or information intelligence applications. They can be used on the one hand to frighten or warn, or on the other hand to signal defeat or flight.

2.5. Copulation Behavior Displays

There are displays indicating *copulation behavior*:

> *"Some displays are performed only before or during the social interactions in which eggs are fertilized. These interactions involve either copulation or some behavioral analogue such as the amplexus behavior of frogs" [60, p. 97].*

Possible function in agent communities. This class of social display would seem to have little to do with agents—at least at their current stage of development. However, analogues to these displays may be pertinent when certain forms of intricate inter-coordination are occurring among agents, involving the need for complex cooperation and coordination to carry out the task successfully, e.g., exchanging ontologies. The copulatory displays are, after all, cues to the parties involved in complex, interdependent operations designed to get an important job done. In a simple fashion, a Palm PDA demonstrates this kind of display when it beeps and lights up after successful docking in its cradle.

2.6. Association Maintenance Displays

There are displays associated with *maintaining, staying-in* association:

> *"Some displays correlate with the behavior involved in remaining with another individual. When individuals so associate they remain together because one, both, or all will follow, will not leave when the other may not follow, and because each permits the others to be nearby…. These displays are not common when animals can maintain their association with ease, but are used primarily when other behavior may disrupt the group. For instance, disruption may result when an individual has just attacked a companion, or flees from an approaching predator before the rest of the group reacts, or even when an individual that has been absent approaches to resume peaceful associating with the group… or when an individual [is] about to move some distance from its group in seeking another foraging site, or by an animal able to maintain contact with its associates only auditorily" [60, p. 104].*

These displays appear to provide a kind of reassurance to other group members that, despite some possible indications to the contrary, the individual has not broken ranks with the group. Such assurances are particularly useful when salient events may raise doubt about the continued association. For example, "the likelihood that a group will remain together after one or more have fought with each other or with outsiders can also be increased by displays encoding an association message" [60, p. 104]. Activities, such as foraging or other societal maintenance activities that require an emphasis on individual effort and perhaps separation from the group, are also prominently associated with association displays. For example, mates who are about to be separated for some time exhibit association displays upon leaving and maintain these messages during the period of separation to the extent possible (e.g. by special vocalizations or gestures—”kissing good-bye and calling home every night,” so to speak) [60, p. 104].

Example forms. Special (often oblique) body orientations toward group members. Various kinds of vocalizations—clearly, signals that can operate over a distance are important in this function.

Possible functions in agent communities. Ways of indicating allegiance to the team and team goals would seem to have a useful place in agent groups and teamwork, especially when some agent is temporarily stepping beyond normal bounds of location or activity for whatever reason. When agents exercise physical mobility, the reporting of location, continued association, and continued commitment to group intent would seem to have a potentially beneficial role.

2.7. Indecisiveness Displays

Displays indicating *indecisiveness* signal that the individual is in a state of indecision about what to do next. Indicators of indecision are various, ranging from simply adopting a static, frozen stance, as if waiting for the situation to provide greater cues, to variations on displays that usually indicate action but are modified to increase the range of choice. An example of the latter would be moving back-and-forth laterally (“pacing”) with respect to a pertinent stimulus, as opposed to approaching it or backing away. Displays for indecisiveness can include behaviors irrelevant and inappropriate to the situation, e.g., suddenly, unexpectantly initiating grooming or eating [60, p. 107].

Example forms. Irrelevant behavior. Moving back-and-forth laterally in relation to a stimulus.

Possible functions in agent communities. It may be useful for an agent to signal that it does not know what to make of some situation, that it is “confused,” or cannot figure out what to do next as a means for eliciting help from humans or other agents.

2.8. Locomotion Displays

Displays indicating *locomotion* simply signal that the animal is moving or is about to move:

> *“[These] displays provide information about a communicator’s use of flight (or other locomotive) behavior, but* not *about functional categories of flight such as approach, withdrawal, attack, or foraging. The displays correlate with all these acts and more…some [animals] extend the performance of the displays to correlate with hopping or running when they forage on the ground. Thus the behavior is viewed as ‘locomoting’ rather than as ‘flying…’” [60, p. 108].*

Example forms. These displays appear to consist primarily of various forms of vocalizations. However, signals indicating that an animal is *about* to move can be more diverse, for example, dances in honeybees, head-tossing in geese.
Possible functions in agent communities. Signals that indicate that an agent is moving or is about to move would seem particularly germane in teams containing mobile agents. As an example of such a display as a warning, think of the distinctive sound that large trucks make when they are about to move in reverse.

2.9. Staying-Put Displays

Displays indicating *remaining with a site* are the opposite of the locomotion displays:

> *"Displays performed only when a communicator is remaining at a fixed site encode the information he will remain at a single point, in the vicinity of such a locus, or in an area that allows considerable movement within fixed boundaries. The behavioral selection referred to is simply "staying-put," defined with respect to a site" [60, p. 115].*

Example forms. Song vocalizations, in particular, are associated with remaining in a territory. Birds that do not sing can have special vocalizations for remaining in place, e.g., the "ecstatic" vocalization of the Adelie penguin [60, p. 115]. Also included are wing-beating, and various specialized postures and movements.
Possible functions in agent communities. As with displays of locomotion, displays of "staying put" are pertinent to mobile agents.

2.10. Attentiveness Displays

Displays indicating *attentiveness to a stimulus* simply convey that the communicator is attending to something and monitoring it.
Example forms. Three distinct barks of a prairie-dog, indicating three different phases of monitoring. Such barks might indicate, for example, that a predator is in the vicinity.
Possible functions in agent communities. For agents, these signals could portend that something important might be happening. It would be useful in agent communities to have a general indicator of alert, that something significant might be transpiring at a particular location or involving a particular agent. Appropriate response, of course, would require additional information. In the animal world, for instance, this additional information sometimes indicates the location of the stimulus.

3. Some Sources of Order in the Human World

It is not surprising that joint activity—and the "getting to know each other" both necessary for it and engendered by it—are also important to humans. Additionally, many of the same benefits are accrued—in particular, inter-predictability and its relationship to coordination and an orderly society. Moreover, the same basic components are involved, including signals understood by both parties to be an invitation to engage in joint activity. However, because of our wider behavioral repertoire, the greater complexity of our communication processes, and our reduced dependence on biological determinism, human cooperation and regulatory processes take on an even greater variety of forms.[4]

In fact, *because* of our vast behavioral repertoire, and because we are so underdetermined in our biology, the argument has been made that a very large portion of what humans do and create is constituted to "control ourselves"! In this view, the role of human culture is that of a vast, fabricated self-regulatory mechanism [29]:[5]

> *I want to propose two ideas: The first of these is that culture is best seen not as complexes of concrete behavior patterns—customs, usages, traditions, habit clusters—as has, by and large, been the case up to now, but as a set of control mechanisms—plans, recipes, rules, instructions (what computer engineers call 'programs')—for the governing of behavior. The second idea is that man is precisely the animal most desperately dependent upon such extragenetic, outside-the-skin mechanisms, such cultural programs, for ordering his behavior...* (p. 44)
>
> *Man is in need of such symbolic sources of illumination [i.e., human-created cultural control mechanisms—addition ours] to find his bearings in the world because the non-symbolic sort that are constitutionally engrained in his body cast such a diffuse light. The behavior patterns of lower animals are, at least to much greater extent, given to them with their physical structure: genetic sources or information order their actions within much narrower ranges of variation, the narrower and more thouroughgoing the lower the animal. For man, what are innately given are extremely general response capacities, which, although they make possible far greater plasticity, complexity, and, on the scattered occasions when everything works as it should, effectiveness of behavior, leave it much less precisely regulated. This then, is the second face of our argument: Unregulated by cultural patterns—organized systems of significant symbols—man's behavior would be virtually ungovernable, a mere chaos of pointless acts and exploding emotions, his experience virtually shapeless. Culture, the accumulated totality of such patterns, is not just an ornament of human existence but—the principal basis of its specificity—an essential condition for it.* (pp. 45-46)

In summary, according to this argument people create and have created cultures and social conventions—albeit in many disparate forms across mankind that can be hard for outsiders to understand—to provide order and predictability. This is also the main reason we claim, following Smith's arguments above that animals cooperate at all, when they do—that is, in order to make themselves better known and more

[4] For a comprehensive and interesting treatment of these kinds of issues regarding joint activity in humans, see[15].

[5] We recognize that Geertz represents only one of many views of culture, but a discussion of competing views is beyond the scope of this paper.

predictable to each other. Furthermore, it would seem to follow from Geertz's argument that the more autonomous the agents involved, the more need there is for such regulation and the wider the variety of forms it might take.

Order and predictability may have a basis in the simple cooperative act between two people, in which the parties "contract" to engage together in a set of interlinked, mutually beneficial activities. From this simple base, in humans at least, there are constructed elaborate and intricate systems of regulatory tools, from formal legal systems, to standards of professional practice, to norms of proper everyday behavior (along with associated methods of punishment or even simple forms of shaming for violations of these).

4. The Problem of Adaptability

While the discussion so far has dealt mainly with the maintenance of order, change is also necessary in perpetuating healthy societies, especially if those societies are expected to adapt to new circumstances and endure over long periods of time. While we cannot investigate adaptation mechanisms in depth in this paper, we feel it important to point out that such mechanisms of change are recognized as critical in both animal and human societies.

For instance, while animal and human signals carry a certain core nature and meaning in a given community, this meaning is not completely rigid or mechanical, and may be very different in different contexts.[6] Such interaction can often best be described as a kind of "improvisation"—embodying considerable novelty while respecting the rules of the form [53].

Take for example two different cases of a human signaling a call to joint activity with another, in fact, signaling the *same* call in the two cases, "help me." In the first instance, the solicitor is sinking in quicksand, and in the other case the solicitor is posed at one end of a heavy table that needs moving. The particulars of what might ensue will depend on the nature of the two different circumstances but also on the particular individuals involved. In the first case the party responding to the request for help may try to throw a rope. However, if there were a history of bad will between himself and the person in the quicksand, he might also just lay back and watch him slowly sink [lack of will]. In the second case, the party responding to the request for help might, on the one hand, go to the unmanned end of the table and try to help lift (and he would not throw a rope—due to the basic circumstantial difference). On the other hand, he might not—his response may depend on how strong he thinks he is or if he has sustained an injury (degree of capability) or depending on his personal

[6] Norman [50, p. 130] gives the following example of this phenomenon from traffic behavior in different countries: "In Mexico, one wins by aggression. In Britain, one wins by politeness and consideration. [In Mexico, when two cars approach a narrow bridge from different directions, flashing your headlights means, 'I got here first, so keep out of my way.' However] in Britain, in a similar situation, the car that flashes its lights first is signaling, 'I see you, please go ahead and I will wait.' Imagine what happens when a Mexican driver encounters a British driver."

history of experienced helpfulness from the individual making the request [(lack of will)—"He never helps me!"] [after[60, p. 224]].

Thus the elements of consistency, but also potential novelty, may both be necessary to signaling activity in the real world, because the world is never static:

> *"In all social events, the behavior of participants must engender considerable predictability. Without predictability, events falter and their orderliness dissipates… [But] the dilemma addressed in this volume is that development of shared signals and codes necessarily leads to conformity in signaling, but conformity cannot cope well with changing or novel events and, when rigid, is maladaptive"* [61, p. 366].

Hence, all signaling must accommodate elements of variation in the pertinent core joint activity conveyed by the signal. These variations are sensitive at least to the particulars of the circumstances and parties involved; that is, the variation on the core activity is context-sensitive:

> *"Another crucially important aspect of all communication is that it is context-dependent. That is, although the information made available by a formalized signal is largely consistent, the significance or 'meaning' of that information in any event, its interpretation by the individual responding to the signal is affected by information in addition to the signal. This is another form of openness in communication and an important means of dealing with novelty. This requisite ability to alter responses to signals as circumstances change is also the basis of our calibration of individual signalers. Clues to their identities become clues to the specific significances of their signals, although only for individuals who are sufficiently familiar with them"* [61, p. 368].

With regard to change and adaptation in *culture* and its regulatory role, modern biologists have increasingly emphasized that the natural selection process includes not only basic biology but also the equally complex elements of culture, cultural change and cultural selection. For instance Mayr has emphasized that "a person is a target of selection in three different contexts: as an individual, as a member of a family… and as a member of a social group." [44, p. 251]. The latter two, at least, implicate culture in the sense we have been addressing it in this essay. Geertz has gone farther, essentially arguing that humanity and culture are so tightly intertwined that the human-culture system is the unit of selection [29, p. 67]. In short, in enduring societies, culture is not static.

Although such nuanced tailoring of communication and culture to circumstance may not always prove necessary in the working interactions of pure agent teams, the need for such tailoring and adjustment will almost surely arise in mixed human-agent teams, as their work together becomes increasingly consequential and as they sustain their interactions for long periods. This is another key element of making agents acceptable to humans. To be acceptable to humans, agents must conform to certain standards of predictability, but they also must not exhibit bald, naïve-looking rigidity.

While recognizing the importance of adaptation, because of the tremendous challenges currently involved in machine learning, our own work has been initially focused on understanding and enabling various forms of order in agent communities. We will briefly address adaptation again in section 6.

5. Building Cultures for Agent Communities: Sources of Order

Our agent research and development efforts over the past decade have maintained a consistent trend. We have been progressively off-loading selected classes of knowledge, some aspects of decision-making, and various kinds of specialized reasoning and problem solving from individual agents into a common environment shared by all agents of a given community, regardless of the nature or sophistication of their internals or the platform on which they are running.[7] This has taken the form, for instance, of the creation of various types of services and various bodies of policy that help regulate conduct across communities of heterogeneous agents running on various platforms. It is in this sense that what we have been doing might be thought of as creating "cultures" for agent communities, especially communities that might endure for long periods of time. We have termed this kind of approach "terraforming cyberspace" (referring to the aspect of the effort that aims to make networked environments a more habitable place for agents) and "cyberforming terraspace" (referring to the aspect of the effort that aims to embed socially-competent agents in the physical world) [12].

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. 1. Required elements of future infrastructure for agents

To support sustainability of groups of agents over long periods, we have envisioned basic types of services that will be needed (figure 1). At a minimum, future infrastructure must go beyond the bare essentials of support to provide pervasive *life support services* (relying on mechanisms such as orthogonal persistence [36] and strong mobility [62; 63]) that help ensure the survival of agents designed to live for long periods of time. Beyond the basics of individual agent protection, these communities will depend on *legal services*, based on explicit policies, to ensure that rights and obligations are monitored and enforced. Benevolent *social services* might also be provided to proactively avoid problems and help agents fulfill their obligations. Although some of these elements exist in embryo within specific agent systems, their scope and effectiveness has been limited by the lack of underlying support at both the platform and application levels.

[7] It could also be said that we have been moving elements from the "sharp end" to the "blunt end" of agents' activity, as these two terms have been characterized by David Woods and colleagues [20].

In the remainder of this section, we will briefly review efforts to create and regulate agent cultures through the use of norms and policies (5.1)[8]. We will discuss the relationship between plans and policy (5.2) and between autonomy and policy (5.3). We will introduce KAoS (5.4) and some basic categories of technical and social policies (5.5). Then we will provide a few examples of policies that address joint activity and signaling, that we are developing for military and space applications (5.6).

5.1. Norms and Policy

In the early 20th century, a legal theorist named Wesley Newcomb Hohfeld developed a theory of fundamental legal concepts [32] from which most of current work on theories of normative positions have taken at least some degree of inspiration (see e.g., [40; 57]).

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 [3]. 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 transgression. 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).[9]

Shoham and Tennenholtz [58] 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 [71], *norm-based* approaches have grown in popularity [6; 21; 41; 42]. In the multi-agent system research community, Conte and Castelfranchi [19] found that norms were variously described as constraints on

[8] We have concentrated first on mechanisms for establishing order and predictability in agent communities because at the current state of agent development these seem to be the greatest concerns of both producers and consumers of agent technologies. Others have focused on issues of "democracy," micro-economics, and other forms of relative freedom in open societies of agents e.g. [14][45][49].

[9] In an insightful essay, Roger Clarke explores some of the implications of Asimov's stories about the laws of robotics for information technologists [16]. Weld and Etzioni [72] 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 [52] 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, which are motivated to cooperate and follow optimally generated policies.

behavior, ends or goals, or obligations. For the most part, implementations of norms in multi-agent systems share three basic features:

- they are designed offline; or
- they are learned, adopted, and refined through the purposeful deliberation of each agent; and
- 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) [22; 37; 67]. 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, 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 explicitly recognized authority.[10] 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:

- support dynamic runtime policy changes, and not merely static configurations determined in advance;
- 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 agents can comply with policy; and
- wherever possible they are enforced preemptively, preventing buggy or malicious agents from doing harm in advance rather than rewarding them or imposing sanctions on them after the fact.

5.2. Plans and Policy

Policy management 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) [27, pp. 161-162]. 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

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

planning. The independence of policy, reasoning, and enforcement mechanisms from planning capabilities helps assure that, wherever possible, key constraints imposed by the humans are respected even in the face of buggy or malicious agents on the one hand, and poorly designed or oversimplified plans on the other. 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."[11]

5.3. Autonomy and Policy[12]

Some important dimensions of the relationship between autonomy and policy can be straightforwardly characterized by reference to figure 1. [13]

The outermost rectangle, labeled *potential actions,* represents the set of all actions defined in some ontology under current consideration.[14] In other words, it contains the union of all actions for all actors currently known to the computational entities that are performing reasoning about adjustable autonomy and mixed-initiative interaction. Note that there is no requirement that all actions that an agent 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 achievement by some agent is deemed sufficiently imaginable in the current context. Of these possible actions, any given actor[15] (e.g., Agent A) will likely only be deemed to be *capable of* performing some subset. Capability is a function of the *abilities* and *resources* available to an actor attempting to undertake some action. An actor's ability is the sum of its own knowledge and skills, whereas its resources consist of all other assets it can currently draw on in the performance of the action. Two actors, Agent A

[11] We are exploring the relationship between policy and planning in new research with James Allen [2][9].

[12] More detail on this topic can be found in [9].

[13] We can make a rough comparison between some of these dimensions and the aspects of autonomy described by Falcone and Castelfranchi [25]. 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.

[14] 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.

[15] 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).

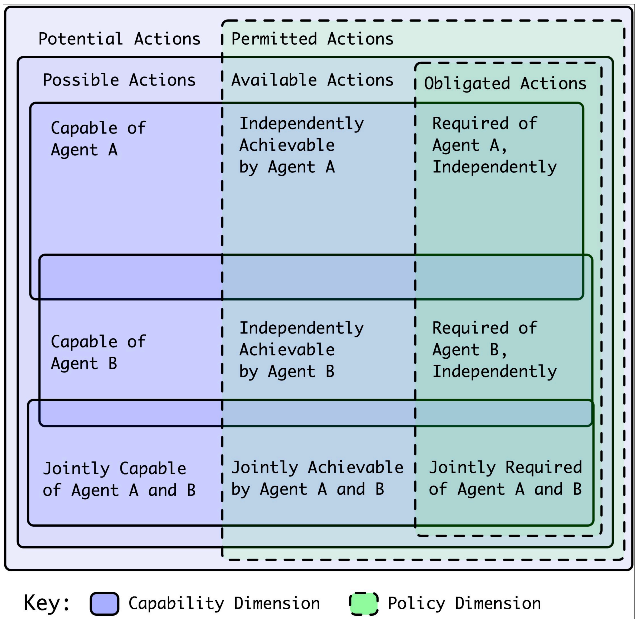
Potential Actions
Permitted Actions
Possible Actions
Available Actions
Obligated Actions
Capable of Agent A
Independently Achievable by Agent A
Required of Agent A, Independently
Capable of Agent B
Independently Achievable by Agent B
Required of Agent B, Independently
Jointly Capable of Agent A and B
Jointly Achievable by Agent A and B
Jointly Required of Agent A and B
Key:
Capability Dimension
Policy Dimension

Fig. 2. Basic dimensions of adjustable autonomy and mixed-initiative interaction.

and Agent B, may have both overlapping and unique capabilities.[16] If a set of actors is *jointly capable of* performing some action, it means that it is deemed to be possible for it to be performed by relying on the capabilities of both actors. Some actors may be capable of performing a given action either individually or jointly; other actors may not be so capable.

In addition to the *descriptive* axis describing various dimensions of capability, there is a *prescriptive* axis that is defined by policies specifying the various *permissions* and *obligations* of actors. *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 so long as the agreement is in effect. The *effectivity* of an individual policy is the set of conditions that determine when it is in or out of force.

The set of *permitted actions* is defined by *authorization policies* that specify which actions an actor is allowed (*positive authorizations* or *A+* policies) or not allowed (*negative authorizations* or *A-* policies) to perform in a given context. The intersection of what is possible and what is permitted to a given set of actors defines a set of *available actions*.

Of those actions that are available to a given actor, some subset may be judged to be *independently achievable* by it in the current context. Some actions, on the other hand, would only be *jointly achievable*.

[16] Note that although we show A and B sharing the same set of possible actions in figure 2, this is not necessarily the case.

Finally, the set of *obligated actions* is defined by *obligation policies* that specify actions that an actor is required to perform (*positive obligations* or *O+* policies) or for which such a requirement is waived (*negative obligations* or *O-* policies). Positive obligations commit the resources of actors, reducing their current overall capability accordingly. *Jointly obligated actions* are those that two or more agents are explicitly required to perform.

A major challenge is to ensure that the degree of autonomy is continuously and transparently adjusted to be consistent with explicitly declared policies which themselves can, ideally, be imposed and removed at any time as appropriate [48]. For example, one goal of the agent or external entity performing such adjustments should be to make sure that the range of permissible actions do not exceed the range of those that are likely to be achievable by the agent.[17] While the agent is constrained to operate within whatever deontic bounds on autonomy are currently enforced as authorization and obligation policies, it is otherwise free to act.

Thus, the coupling of autonomy with policy gives the agent maximum opportunity for local adaptation to unforeseen problems and opportunities, while assuring humans that agent behavior will be kept within desired bounds.

In principle, the actual adjustment of an agent's level of autonomy could be initiated either by a human, the agent, or some other software component.[18] 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 renegotiate roles and tasks among humans and agents as the situation demands. Such adjustments can also be anticipatory when agents are capable of predicting the relevant events [5; 25]. Research in adaptive function allocation—the dynamic assignment of tasks among

[17] If the range of achievable actions for an agent is found to be too restricted, it can, in principle, be increased in any combination of four ways: 1. removal of some portion of the environmental constraints, thus increasing the range of possible actions; 2. increasing its permissions; 3. making additional external help available to the agent, thus increasing its joint capabilities; or 4. reducing an agent's current set of obligations, thus freeing resources for other tasks. Of course, there is a cost in computational complexity to increasing the range of actions that must be considered by an agent—hence the judicious use of policy where certain actions can either be precluded from consideration or obligated with confidence in advance by a third party.

[18] Cohen [18] draws a line between those approaches in which the agent itself wholly determines the mode of interaction with humans (mixed-initiative) and those where this determination is imposed externally (adjustable autonomy). Additionally, mixed-initiative systems are considered by Cohen 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 the human and also of 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 those systems that are capable of making context-appropriate adjustments to their level of social autonomy (i.e., their level or mode of engagement with the human), whether a given adjustment is made as a result of reasoning internal to the agent or due to externally-imposed policy-based constraints.

humans and machines—provides some useful lessons for implementations of adjustable autonomy in intelligent systems [31].

When evaluating options for adaptively reallocating tasks among team members, it must be remembered that dynamic role adjustment comes at a cost. Measures of expected utility can be used to evaluate the tradeoffs involved in potentially interrupting the ongoing activities of agents and humans in such situations, in order to communicate, coordinate, and reallocate responsibilities [18; 33; 34]. 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 shift in roles among a human-agent pair, but rather the distribution of dynamic demands across many coordinated actors.[19] Defining explicit policies for the transfer of control among team members and for the resultant required modifications to coordination constraints can prove useful in managing such complexity [54]. Whereas in the past goal adoption and the commitment to join and interact in a prescribed manner with a team have sometimes occurred as part of a single act in early teamwork formulations, researchers are increasingly realizing the advantages of allowing the respective 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 [4; 48]. Over the last few years, we have been developing a set of services within a framework called KAoS to accomplish just these kind of goals.

5.4. Overview of KAoS

KAoS is a collection of componentized agent services compatible with several popular agent frameworks, including Nomads [63], the DARPA CoABS Grid [38], the DARPA ALP/Ultra*Log Cougaar framework (http://www.cougaar.net), CORBA (http://www.omg.org), and Voyager (http://www.recursionsw.com/osi.asp). While initially oriented to the dynamic and complex requirements of software and robotic 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 [35].

KAoS domain services provide the capability for groups of agents, people, resources, and other entities to be structured into organizations of agent domains and subdomains to facilitate agent-agent collaboration and external policy administration.

KAoS policy services allow for the specification, management, conflict resolution, and enforcement of policies within domains. The KAoS Policy Ontologies (KPO), represented in OWL [69], distinguishes between *authorizations* (i.e., constraints that

[19] As Hancock and Scallen [31] rightfully observe, the problem of adaptive function allocation is not merely one of efficiency or technical elegance. Economic factors (e.g., can the task be more inexpensively performed by humans, agents, or some combination?), political and cultural factors (e.g., is it acceptable for agents to perform tasks traditionally assigned to humans?), or personal and moral considerations (e.g., is a given task enjoyable and challenging vs. boring and mind-numbing for the human?) are also essential considerations.

permit or forbid some action) and *obligations* (i.e., constraints that require some action to be performed, or else serve to waive such a requirement) [22].

5.5. Technical and Social Policy Categories

To increase the likelihood of human acceptance of agent technology, successful systems must attend to both the technical and social aspects of policy [51]. From a *technical perspective*, we want to be able to help ensure the protection of agent state, the viability of agent communities, and the reliability of the resources on which they depend [12]. 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 in both human and computational environments. From a *social perspective*, we want agents to be designed to fit well with how people actually work together and otherwise interact. Explicit policies governing human-agent interaction, based on careful observation of work practice and an understanding of current social science research, can help assure that effective and natural coordination, appropriate levels and modalities of feedback, and adequate predictability and responsiveness to human control are maintained [11; 26; 50]. These and similar technical and social factors are key to providing the reassurance and trust that are the prerequisites to the widespread acceptance of agent technology for non-trivial applications.

We currently classify *technical policies* into six categories:

- *Authentication policies*. This category of policies is concerned with assuring that identification of proper users is associated with various agent commands and actions.
- *Data and resource access and protection policies*. These policies control access to resources, information, and services, and specify any constraints on data protection (e.g., encryption).
- *Communication policies*. Communication policies govern message passing among individuals and groups, including forms of content filtering and transformation [64; 65].
- *Resource control policies*. Going beyond simple access control, these policies control the amount and rate of resource usage (e.g., CPU, memory, network, hard disk, screen space) [62; 63].
- *Monitoring and response policies*. These policies typically represent obligations for the system to perform specific monitoring and response actions (e.g., logging, response to authorization failures or changes to global system defense postures).
- *Mobility policies*. Mobility policies govern the physical movement of software or hardware agents [39].

We are also currently developing *social policies* within six categories:[20]

[20] The motivation for such policies is eloquently expressed by Norman [50, p. 126-127]: "One of the reasons that modern technology is so difficult to use is because of [its] silent, invisible operation [when compared with mechanical devices]. The videocassette recorder, the digital watch, and the microwave oven—none is inherently complicated. The problem for us is their lack of communication. They fail to interact gracefully. They demand attention and services,

- *Organization policies.* This category of policies includes those that specify relationships among classes of agents, e.g., policies about delegation of responsibilities and agent registration in domains.
- *Notification policies.* It is important that important information be conveyed to the appropriate people at the appropriate time and with an appropriate modality. Based on the work of [55; 56], we are building ontologies supporting policies for categories of agents, roles, notifications, latency, focus of attention, and presence as a foundation for policies governing context-sensitive notification.[21]
- *Conversation policies.* Explicit conversation policies simplify the work of both the agent and the agent designer [30]. Such policies include, for example, constraints on conversation message sequencing, termination conditions, and timeouts.
- *Nonverbal expression.* These policies govern signaling and display behavior of agents. Detailed examples are given below.
- *Collaboration policies.* Policies governing team coordination are classed in this category, including the formation and discharge of joint goals as is central in traditional multi-agent teamwork theory [17; 66].
- *Adjustable autonomy policies.* These policies regulate levels of, and adjustments to, levels of agent autonomy [23].

A fuller discussion of examples from each of these categories may be found in [7].

We now discuss a few simple examples of policy relating to the theme of display and signaling behavior. The policy examples are drawn from our studies of the Personal Satellite Assistant (PSA), currently under development at NASA Ames. The PSA is a softball-sized flying robot, designed to help astronauts, that is being developed to operate onboard spacecraft in pressurized micro-gravity environments [1; 10; 24; 28; 59].

For clarity's sake, we will present example policies in ordinary English rather than in DAML. For brevity's sake, the policies will be presented in an incomplete form. Each example is preceded by A+, A-, O+, or O- to indicate whether it is respectively a positive authorization, negative authorization, positive obligation, or negative obligation.

but without reciprocating, without providing sufficient background and context. There is little or no feedback.… The modern information-processing machine fits the stereotype of an antisocial, technological nerd. It works efficiently, quietly, and autonomously, and it prefers to avoid [social] interactions with the people around it."

[21] Of course the important point in context-sensitive notification in our day of information and sensory overload is sometimes not helping the information get through but rather blocking, filtering, or rechanneling it in appropriate ways: "Most instrument panels [use lights, buzzers, and alarms to] tell us when something is wrong and needs immediate attention. No social protocols, no etiquette. No checking to see whether we are busy at some other activity, usually not even a check to see if other alarms or warnings are also active. As a result, all the alarms and warnings scream in their self-centered way… In places that have large control panels,… the first act of the human operators is to shut off the alarms so they can concentrate upon the problem" [50, p. 128].

5.6. Nonverbal Expression Policy: Examples

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 [47, p. 295]. On the other hand, animals and humans often rely on visual and auditory signals in place 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 robotic and software agents. These nonverbal behaviors will be designed to express not only the current state of the agent but also—importantly—to provide rough clues about what it is going to do next.

Maes and her colleagues were among the first to explore this possibility in her research on software agents that continuously communicated their internal state to the user via facial expressions (e.g., thinking, working, suggestion, unsure, pleased, and confused) [43]. Breazeal has taken inspiration from research in child psychology [68] to develop robot displays that reflect four basic classes of preverbal social responses: affective (changing facial expressions), exploratory (visual search, maintenance of mutual regard with human), protective (turning head away), and regulatory (expressive feedback to gain caregiver attention, cyclic waxing and waning of internal states, habituation, and signals of internal motivation) [13]. Norman has investigated the role of signaling, not only in effective coordination and communication, including the communication of emotion, but also with regard to the role of deception in human and agent affairs [50]. Books on human etiquette [70] contain many descriptions of appropriate behavior in a wide variety of social settings. Finally, in addition to this previous work, we think that display and signaling behavior among people [46] 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 ten categories of widespread vertebrate animal cooperation and coordination displays [60, pp. 84-105]. Some examples are discussed below.

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

This policy would require 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 tasks 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 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, so long as they are within the line of
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 a person know that it is still attending to the ongoing conversation and/or task.

```
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 it's got to be on the move. This policy prevents the PSA from moving until it has first signaled for some number of seconds its intention to move. Besides the pre-move signaling, some kind of signaling could also take place during the move itself.

In addition to this policy regarding movement, other policies should be put in place to, for instance, require the PSA to stay at a safe and comfortable distance from people, other robotic agents, and space station structures and equipment. Of course the policies would take into account that different social distances may be appropriate in different cultures, as will be pertinent in, for example, multinational operations of the International Space Station.

As our new phases of research proceed, we hope to verify the effectiveness of KAoS policies and services through a series of tests assessing *survivability* (ability to maintain effectiveness in the face of unforeseen software or hardware failures), *safety* (ability to prevent certain classes of dangerous actions or situations), *predictability* (assessed correlation between human judgment of predicted vs. actual behavior), *controllability*(immediacy with which an authorized human can prevent, stop, enable, or initiate agent actions), *effectiveness* (assessed correlation between human judgment of desired vs. actual behavior), and *adaptability* (ability to respond to changes in context). We briefly address some aspects of adaptation next.

6. Building Cultures for Agent Communities: Potential Sources of Adaptation

There are two sorts of adaptation we believe will be critical to capture if communities of agents are to be enduring. The first has to do with the adaptation of policy to accommodate diverse contexts over which it must be applied. For example, we have seen an example of the need for this kind of adaptation in the last section, in which the comfortable distance a PSA should keep from its partner invokes cultural considerations.

The limited progress we have made with regard to adaptation to context has been mostly in the area of adjustable autonomy and the capability it provides for functions like dynamic handoff of control among team members and flexible renegotiation of roles and tasks among humans and agents when new opportunities arise or when breakdowns occur [see section 5.3 and [9]].

The second type of adaptation involves changes in policy, either in response to experience, for example, in realizing that enforcing a policy or set of policies has consistently resulted in untoward outcomes, or by recognizing that the nature of the operational world had changed in consequential ways. This second kind of adaptation has been even less explored. From the perspective of this paper, such adaptation might involve a sort of "cultural learning" that might prove challenging to current machine learning approaches.

7. Conclusion

In this paper, we have attempted to encourage an expansion of thinking about the sources, nature, and diversity of regulatory systems that can be utilized to achieve acceptable levels of order when groups of agents or mixed agent-human groups are engaged in consequential work. Interestingly, one impetus for this direction has been a desire to "make agents acceptable to humans," for example, to make communication with agents natural, to make agents seem trustworthy (and actually be trustable) in their participation in important affairs, and perhaps most importantly, to ensure, as in human societies, a kind of predictability—agents will not be acceptable to humans if they unexpectedly keep running amok.

In addition, recognizing that societies need to adapt to changing conditions in addition to maintaining order, we have examined elements of adaptation in animal and human groups. Since healthy order can lapse into ineffective and unacceptable rigidity, we have made some brief speculations about ways elements of useful adaptation might coexist with those enforcing order.

While we have focused primarily on animal signaling and order, we anticipate, especially in situations in which agents are embodied (e.g., physical robots), and move around, and act in the world, that there will be considerable benefit from expanding our repertoire of agent-cultural devices even farther, to include, for example, concrete instantiations such as "lines on the highway" or more subtle codes of "good manners."

Some years ago, Paul Wohlmuth, a philosopher of law, wrote the introductory chapter to an interdisciplinary special issue of the *Journal of Contemporary Legal Issues* focused on the "constitution of authority" [73]. Roughly interpreted, the constitution of authority refers to how things of various sorts come to have regulatory power over human conduct.

In his introductory chapter, Wohlmuth used the simple example of an automobile navigating a bend in the road[22] to illustrate the ubiquity and wide variety of authoritative forms that come to bear on human activity. Even the basic laws of physics are involved. That is, there are limits to the speed with which a particular sort of car, on a particular sort of road, can navigate the turn without crashing, and people who do not want to get hurt will honor these constraints as they are able. The laws of basic physiology are in place, for instance, the eyesight, reaction time, and degree of alertness of the driver. Beyond these more physical constraints, all sorts of cultural artifacts are imparting regulatory influence. Stripes on the road demarcate the lanes and boundaries of the highway and whether or not the lanes may be traversed. Regional custom determines which side of the highway the driver should be on at all. Signs containing both words (e.g., “slow down”) and symbols (e.g., a twisting portrayal of the road section) are present. There are also inter-vehicular signalings of intent and disposition, and processes of coordination taking place, if there are multiple vehicles present. Furthermore, the appearance of a law-enforcement official, for example a patrolling police car, emerging on the scene, has dramatic effect on the behavior of the drivers. At much greater degrees of abstraction from the scene, there is the Motor Vehicle Code and other formal statutes that, for instance, prescribe the amount of certain substances that the driver may have in his or her body. In addition, there are entire culturally constructed deliberative bodies (e.g., the courts) empowered, if needed, to bridge the gap between pertinent statutes and the particulars of any one instance of traveling this bend. And, not much, if any, of this is static. For instance, if a particularly high rate of accidents results at this bend, many changes may take place, ranging from the physical to the more abstract. The road banking on the curve may be increased. The posted speed limit may be decreased. More ominous, scarier symbols may be posted. Consequences of violations of pertinent rules of the road may be made harsher.

Societies must maintain a degree of continuity and stability; this represents a kind of historical memory of practices that have been effective in the past and have supported survival. On the other hand, the world is ever-changing; continued survival requires adaptations in practice to address novelty and surprise. The complexity of this interplay makes us realize even more that we are only at the beginning in addressing the dual problems of order and change in agent communities (let alone the optimal delicate balance between them), and it is hard not to feel a bit overwhelmed. But we are convinced that even little steps in understanding how to better incorporate the content and mechanisms of culture into agent societies will be both interesting and fruitful.

8. Acknowledgments

The authors gratefully acknowledge the sponsorship of this research by the NASA Cross-Enterprise and Intelligent Systems Programs, and a joint NASA-DARPA ITAC

[22] Interestingly, Smith [60] and Norman [50] also draw extensively on analogies to highway traffic to illustrate their discussion of signaling and coordination.

grant. Additional support was provided by DARPA's CoABS, Ultra*Log, EPCA, and Augmented Cognition programs, by the Army Research Lab's Advanced Decision Architectures program (ARLADA), and the Office of Naval Research. A special thanks is due to the following individuals for their valuable contributions to the ideas developed in this paper: Alessandro Acquisti, James Allen, Guy Boy, Kathleen Bradshaw, Maggie Breedy, Larry Bunch, Murray Burke, Nate Chambers, Bill Clancey, Greg Dorais, Joan Feltovich, Ken Ford, Lucian Galescu, Mark Greaves, Jack Hansen, Pat Hayes, Robert Hoffman, Matt Johnson, Hyuckchul, Jung, Shri Kulkarni, Henry Lieberman, Cheryl Martin, Cindy Mason, Robin Murphy, Nicola Muscettola, Don Norman, Jerry Pratt, Anil Raj, Dylan Schmorrow, Debra Schreckenghost, Mike Shafto, Maarten Sierhuis, Austin Tate, Milind Tambe, William Taysom, Ron Van Hoof, the late Paul Wohlmuth, and Tim Wright.

9. References

[1] Acquisti, A., Sierhuis, M., Clancey, W. J., & Bradshaw, J. M. (2002). 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,

[2] Allen, J. F., & Ferguson, G. (2002). Human-machine collaborative planning. *Proceedings of the NASA Planning and Scheduling Workshop*. Houston, TX,

[3] Asimov, I. (1942/1968). Runaround. In I. Asimov (Ed.), *I, Robot*. (pp. 33-51). London, England: Grafton Books. Originally published in *Astounding Science Fiction,* 1942, pp. 94-103.

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

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

[6] Boman, M. (1999). Norms in artificial decision-making. *Artificial Intelligence and Law*, 7, 17-35.

[7] Bradshaw, J. M., Beautement, P., Raj, A., Johnson, M., Kulkarni, S., & Suri, N. (2003). Making agents acceptable to people. In N. Zhong & J. Liu (Ed.), *Intelligent Technologies for Information Analysis: Advances in Agents, Data Mining, and Statistical Learning*. (pp. in press). Berlin: Springer Verlag.

[8] Bradshaw, J. M., Boy, G., Durfee, E., Gruninger, M., Hexmoor, H., Suri, N., Tambe, M., Uschold, M., & Vitek, J. (Ed.). (2003). *Software Agents for the Warfighter. ITAC Consortium Report*. Cambridge, MA: AAAI Press/The MIT Press.

[9] Bradshaw, J. M., Jung, H., Kulkarni, S., & Taysom, W. (2004). Dimensions of adjustable autonomy and mixed-initiative interaction. In M. Klusch, G. Weiss, & M. Rovatsos (Ed.), *Computational Autonomy*. (pp. in press). Berlin, Germany: Springer-Verlag.

[10] Bradshaw, J. M., Sierhuis, M., Acquisti, A., Feltovich, P., Hoffman, R., Jeffers, R., Prescott, D., Suri, N., Uszok, A., & Van Hoof, R. (2003). Adjustable autonomy and human-agent teamwork in practice: An interim report on space applications. In H. Hexmoor, R. Falcone, & C. Castelfranchi (Ed.), *Agent Autonomy*. (pp. 243-280). Kluwer.

[11] Bradshaw, J. M., Sierhuis, M., Gawdiak, Y., Jeffers, R., Suri, N., & Greaves, M. (2001). Adjustable autonomy and teamwork for the Personal Satellite Assistant. *Proceedings of the IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents*. Seattle, WA, USA,

[12] Bradshaw, J. M., Suri, N., Breedy, M. R., Canas, A., Davis, R., Ford, K. M., Hoffman, R., Jeffers, R., Kulkarni, S., Lott, J., Reichherzer, T., & Uszok, A. (2002). Terraforming cyberspace. In D. C. Marinescu & C. Lee (Ed.), *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.

[13] Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. *IROS*. Kyonjiu, Korea,

[14] Calmet, J., Daemi, A., Endsuleit, R., & Mie, T. (2003). A liberal approach to openess in societies of agents. A. Omicini, P. Petta, & J. Pitt (Ed.), *Proceedings of the Fourth Workshop on Engineering Societies for Agents World,* (pp. 13-18). Imperial College London, UK,

[15] Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

[16] Clarke, R. (1993-1994). Asimov's laws of robotics: Implications for information technology, Parts 1 and 2. *IEEE Computer*, December/January, 53-61/57-66.

[17] Cohen, P. R., & Levesque, H. J. (1991). *Teamwork*. Technote 504. Menlo Park, CA: SRI International, March.

[18] Cohen, R., & Fleming, M. (2002). Adjusting the autonomy in mixed-initiative systems by reasoning about interaction. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Ed.), *Agent Autonomy*. (pp. 105-122). Dordrecht, The Netherlands: Kluwer.

[19] Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London, England: UCL Press.

[20] Cook, R. I., & Woods, D. D. (1994). Operating at the sharp end: The complexity of human error. In S. U. Bogner (Ed.), *Human Error in Medicine*. Hillsdale, NJ: Lawrence Erlbaum.

[21] d'Inverno, M., & Luck, M. (2001). *Understanding Agent Systems*. Berlin, Germany: Springer-Verlag.

[22] Damianou, N., Dulay, N., Lupu, E. C., & Sloman, M. S. (2000). *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.

[23] Dorais, G., Bonasso, R. P., Kortenkamp, D., Pell, B., & Schrekenghost, D. (1999). 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,

[24] Dorais, G., Desiano, S. D., Gawdiak, Y., & Nicewarmer, K. (2003). An autonomous control system for an intra-vehicular spacecraft mobile monitor prototype. *Proceedings of the Seventh International Symposium on Artificial Intelligence, Robotics, and Automation in Space (i-SAIRAS 2003)*. Nara, Japan,

[25] Falcone, R., & Castelfranchi, C. (2002). From automaticity to autonomy: The frontier of artificial agents. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Ed.), *Agent Autonomy*. (pp. 79-103). Dordrecht, The Netherlands: Kluwer.

[26] Feltovich, P., Bradshaw, J. M., Jeffers, R., & Uszok, A. (2003). Order and KAoS: Using policy to represent agent cultures. *Proceedings of the AAMAS 03 Workshop on Humans and Multi-Agent Systems*. Melbourne, Australia,

[27] Fox, J., & Das, S. (2000). *Safe and Sound: Artificial Intelligence in Hazardous Applications*. Menlo Park, CA: AAAI Press/The MIT Press.

[28] Gawdiak, Y., Bradshaw, J. M., Williams, B., & Thomas, H. (2000). R2D2 in a softball: The Personal Satellite Assistant. H. Lieberman (Ed.), *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI 2000),* (pp. 125-128). New Orleans, LA, New York: ACM Press,

[29] Geertz, C. (1973). *The Interpretation of Cultures*. New York, NY: Basic Books.

[30] Greaves, M., Holmback, H., & Bradshaw, J. M. (1999). What is a conversation policy? M. Greaves & J. M. Bradshaw (Ed.), *Proceedings of the Autonomous Agents '99 Workshop on Specifying and Implementing Conversation Policies,* (pp. 1-9). Seattle, WA,

[31] Hancock, P. A., & Scallen, S. F. (1998). Allocating functions in human-machine systems. In R. Hoffman, M. F. Sherrick, & J. S. Warm (Ed.), *Viewing Psychology as a Whole*. (pp. 509-540). Washington, D.C.: American Psychological Association.

[32] Hohfeld, W. N. (1913). Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 23.

[33] Horvitz, E. (1999). 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,

[34] Horvitz, E., Jacobs, A., & Hovel, D. (1999). Attention-sensitive alerting. *Proceedings of the Conference on Uncertainty and Artificial Intelligence (UAI '99),* (pp. 305-313). Stockholm, Sweden,

[35] Johnson, M., Chang, P., Jeffers, R., Bradshaw, J. M., Soo, V.-W., Breedy, M. R., Bunch, L., Kulkarni, S., Lott, J., Suri, N., & Uszok, A. (2003). KAoS semantic policy and domain services: An application of DAML to Web services-based grid architectures. *Proceedings of the AAMAS 03 Workshop on Web Services and Agent-Based Engineering*. Melbourne, Australia,

[36] Jordan, M., & Atkinson, M. (1998). *Orthogonal persistence for Java—A mid-term report*. Sun Microsystems Laboratories,

[37] Kagal, L., Finin, T., & Joshi, A. (2003). A policy language for pervasive systems. *Proceedings of the Fourth IEEE International Workshop on Policies for Distributed Systems and Networks,* (pp. http://umbc.edu/~finin/papers/policy03.pdf). Lake Como, Italy,

[38] Kahn, M., & Cicalese, C. (2001). CoABS Grid Scalability Experiments. O. F. Rana (Ed.), *Second International Workshop on Infrastructure for Scalable Multi-Agent Systems at the Fifth International Conference on Autonomous Agents*. Montreal, CA, New York: ACM Press,

[39] Knoll, G., Suri, N., & Bradshaw, J. M. (2001). 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),* (pp. 54-60). Montreal, CA, New York: ACM Press,

[40] Krogh, C., & Herrestad, H. (1999). Hohfeld in Cyberspace and other applications of normative reasoning in agent technology. *Artificial Intelligence and Law*, 7(1), 81-96.

[41] Lopez y Lopez, F., Luck, M., & d'Inverno, M. (2001). A framework for norm-based inter-agent dependence. *Proceedings of the Third Mexican Internation Conference on Computer Science*.

[42] Lopez y Lopez, F., Luck, M., & d'Inverno, M. (2002). Constraining autonomy through norms. *Proceedings of the Conference on Autonomous Agents and Multi-Agent Systems,* (pp. 674-681). Bologna, Italy,

[43] Maes, P. (1997). Agents that reduce work and information overload. In J. M. Bradshaw (Ed.), *Software Agents*. (pp. 145-164). Cambridge, MA: AAAI Press/The MIT Press.

[44] Mayr, E. (1997). *This Is Biology*. Cambridge, MA: Belkamp Press of Harvard University Press.

[45] McBurney, P., & Parsons, S. (2003). Engineering democracy in open agent systems. A. Omicini, P. Petta, & J. Pitt (Ed.), *Proceedings of the Fourth Workshop on Engineering Societies for Agents World,* (pp. 125-131). Imperial College London, UK,

[46] Morris, D. (2002). *Peoplewatching*. London, England: Vintage.

[47] Murphy, R. R. (2000). *Introduction to AI Robotics*. Cambridge, MA: The MIT Press.

[48] Myers, K., & Morley, D. (2003). Directing agents. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Ed.), *Agent Autonomy*. (pp. 143-162). Dordrecht, The Netherlands: Kluwer.

[49] Neville, B., & Pitt, J. (2003). A computational framework for social agents in agent-mediated e-commerce. A. Omicini, P. Petta, & J. Pitt (Ed.), *Proceedings of the Fourth Workshop on Engineering Societies for Agents World,* (pp. 149-156). Imperial College London, UK,

[50] Norman, D. A. (1992). Turn signals are the facial expressions of automobiles. In *Turn Signals Are the Facial Expressions of Automobiles*. (pp. 117-134). Reading, MA: Addison-Wesley.

[51] Norman, D. A. (1997). How might people interact with agents? In J. M. Bradshaw (Ed.), *Software Agents*. (pp. 49-55). Cambridge, MA: The AAAI Press/The MIT Press.

[52] Pynadath, D., & Tambe, M. (2001). Revisiting Asimov's first law: A response to the call to arms. *Proceedings of ATAL 01*.

[53] Sawyer, R. K. (2001). *Creating Conversations: Improvisation in Everyday Discourse*. Cresskill, NJ: Hampton Press.

[54] Scerri, P., Pynadath, D., & Tambe, M. (2002). Adjustable autonomy for the real world. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Ed.), *Agent Autonomy*. (pp. 163-190). Dordrecht, The Netherlands: Kluwer.

[55] Schreckenghost, D., Martin, C., Bonasso, P., Kortenkamp, D., Milam, T., & Thronesbery, C. (2003). Supporting group interaction among humans and autonomous agents. *Submitted for publication*.

[56] Schreckenghost, D., Martin, C., & Thronesbery, C. (2003). Specifying organizational policies and individual preferences for human-software interaction. *Submitted for publication*.

[57] Sergot, M. (2001). A computational theory of normative positions. *ACM Transactions on Computational Logic*, 2(4), 581-622.

[58] Shoham, Y., & Tennenholtz, M. (1992). On the synthesis of useful social laws for artificial agent societies. *Proceedings of the Tenth National Conference on Artificial Intelligence,* (pp. 276-281). San Jose, CA,

[59] Sierhuis, M., Bradshaw, J. M., Acquisti, A., Van Hoof, R., Jeffers, R., & Uszok, A. (2003). 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,

[60] Smith, W. J. (1977). *The Behavior of Communicating*. Cambridge, MA: Harvard University Press.

[61] Smith, W. J. (1995). The biological bases of social attunement. *Journal of Contemporary Legal Issues*, 6.

[62] Suri, N., Bradshaw, J. M., Breedy, M. R., Groth, P. T., Hill, G. A., & Jeffers, R. (2000). 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,

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

[64] Suri, N., Bradshaw, J. M., Burstein, M. H., Uszok, A., Benyo, B., Breedy, M. R., Carvalho, M., Diller, D., Groth, P. T., Jeffers, R., Johnson, M., Kulkarni, S., & Lott, J. (2003). DAML-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,

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

[66] Tambe, M., Shen, W., Mataric, M., Pynadath, D. V., Goldberg, D., Modi, P. J., Qiu, Z., & Salemi, B. (1999). 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,

[67] Tonti, G., Bradshaw, J. M., Jeffers, R., Montanari, R., Suri, N., & Uszok, A. (2003). Semantic Web languages for policy representation and reasoning: A comparison of KAoS, Rei, and Ponder. In D. Fensel, K. Sycara, & J. Mylopoulos (Ed.), *The Semantic Web—ISWC 2003. Proceedings of the Second International Semantic Web Conference, Sanibel Island, Florida, USA, October 2003, LNCS 2870*. (pp. 419-437). Berlin: Springer.

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

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

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

[71] Verhagen, H. (2001). 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,

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

[73] Wohlmuth, P. C. (1995). Traveling the highway: Sources of momentum in behavioral regulation. *Journal of Contemporary Legal Issues*, 6, 1-9.