A Model for Investigating the Effects of Machine Autonomy on Human Behavior

Jeffrey V. Nickerson and Richard R. Reilly
Stevens Institute of Technology
jnickerson@stevens.edu, rreilly@stevens.edu

Abstract

As autonomous machines become more pervasive, situations will arise when human decision-makers will receive advice from both machines and other humans. When these instructions conflict, a new social situation is defined for which we have little precedent. The authors propose a model for investigating these situations. The model synthesizes research from several different fields, including machine autonomy, affect, initial trust, individual differences, and training. The model is explained, and a set of propositions is described. The model is used to analyze the case of an air collision in which machines and humans provided conflicting advice. The model is also applied to situations in which unmanned aerial vehicles and piloted aircraft seek to avoid collisions with each other. Ways of testing the model through human subject experiments are discussed.

1. Introduction

As machines are designed to look and act more like people, it may be that what we currently understand about our interactions with machines is no longer valid. It would be helpful to understand this early, so that we can better design our machines and our training of the users of these machines.

We are proposing a model for research surrounding the effects of machine autonomy on human behavior. We wish to understand how teams that include both autonomous machines and humans interact. We initially focus on a specific problem – how humans respond when confronted with conflicting advice from a machine and another human. In creating this model, we draw from five fields of research. The first is in autonomy. The second is in the human perception of machines, particularly of machines which elicit affect. The third is in the area of trust, specifically initial trust. The fourth is research in individual differences, especially personality. The fifth is situation-specific training. In this paper we synthesize the findings of researchers in these five fields and propose a preliminary model. We then apply the model in two examples from the domain of collision avoidance.

It is clear that the willingness to take advice is related to trust. Trusting a machine, which may mean ceding control to a machine, is related to whether the machine has made errors in the past, as well as to our own self-confidence [1]. If we are overly confident, we will tend to ignore even the good advice of a machine. This issue of delegating to the machine or taking over ourselves has been studied in detail. We bias toward automated aids initially, but in response to errors we tend to distrust and disuse the aids [2]. In team situations, we know that sometimes humans over-promote the machine to the status of a human [3-5].

What appears to be less studied are situations in which the advice of a machine and the advice of a human conflict. We illustrate this in figure 1.

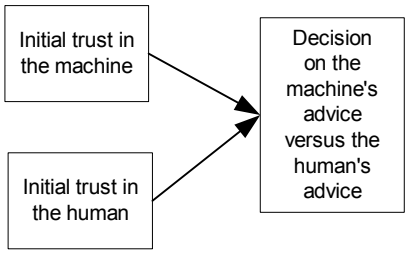

Figure 1.

We think this kind of situation will occur increasingly often as machines increase in autonomy, and enter into teams as active participants.

2. Autonomy

Deciding whether to accept the advice of a machine is influenced by our perception of the machine. While we think it is clear that humans react to machines differently than to humans, we think it is also clear that humans react differently to different types of machines. We wish to understand what characteristics of machines cause different perceptions of the machine. One factor that may affect perceptions is the autonomy of the machine. It may

be that relative autonomy - the ability of the machine to make decisions on its own - changes our perception of the machine. In a rapidly expanding literature, there are many different ways of modeling autonomy [6, 7]. Much of the work in autonomy has roots in the literature of man-machine systems [8-11], including work on the allocation of function [12].

Autonomy can be described along a scale, as in table 1 from Parasuraman, Sheridan and Wickens [13].

Table 1. Levels of autonomy [13].

The computer...	
10	decides everything
9	informs the human only if it wants to
8	informs the human only if asked
7	executes automatically, then informs the human
6	allows the human to veto the machine decision
5	asks for approval
4	suggests one alternative
3	narrows a selection to a few alternatives
2	offers a complete set of alternatives
1	offers no assistance

Hexmoor [14] proposes a specific autonomy metric for collision avoidance. An automated agent will look at trajectories and distances, generating a collision priority between 1 and 4. If collision is imminent, the agent takes over. If collision is a fair distance away, the agent presents a user-interface window to the human pilot on a timer – the human has the length of the timer to respond before the agent takes over. Hexmoor's equation is

$$Autonomy = CollisionPriority/4.0 + ((CollisionPriority - 4.0) * t)/T,$$

where t is the timer, and T is the time to collision.

This metric links autonomy to urgency – in an urgent situation, the agent will take over. If this is implemented, the human will gradually loose the ability to decide as circumstances become more urgent. Inagaki shows that this is optimal in certain situations [15].

There is a problem – as the decision becomes more important, the human has less say. But in some examples, such as in flying a plane, if the automated systems fail, humans need to take over [13]. Any system that lets the skills of the human atrophy through over-automation is dangerous.

Machines with more autonomy may appear more human. Drawing from this observation, we make the following research proposition:

P1: A higher level of autonomy in a machine will increase a human's initial trust in that machine.

It is not obvious that this is true. It essentially says that the less control we have, the greater our trust. We know that, in humans, the amount of autonomy we grant other humans is related to our degree of trust. Ceding control to machines is also related to trust [16]. The reverse proposition, that autonomy influences trust, does not appear to have been studied. It might be argued that a higher level of autonomy in a human will act the same way on trust as higher level of autonomy in a machine, but we are assuming that the effect of human and machine autonomy will differ:

P2: A higher level of autonomy in a machine will affect a human's initial trust to a different degree than a higher level of autonomy in a human will affect another human's initial trust

3. Affect

There is considerable evidence that affect has an important relationship with trust [17]. A recent review by Lee and See argues for the role of affect in trust related to automation [18]. There is evidence that people sometimes treat machines as people [19-21]. In particular, they treat autonomous machines as people. A conversational agent trained to elicit and express affect, although obviously not human, gains more trust through this training, even though participants do not believe the machine itself is experiencing emotion [22]. There is a quickly accumulating body of evidence that, even though we don't appear to be fooled by machines, we respond more positively to those that exhibit human traits [23-28]. Other evidence suggests that synthetic voices that display consistency and are more similar to the respondent will be perceived more favorably and be trusted more [29]. It is possible affect may cause us to overly trust the machine.

In some situations, it appears it is better not to let a machine pretend it is a person. Kiesler, Sproull and Waters [30] showed that, in a prisoner's dilemma, people generally played fairly with a computer – but that a text interface worked better than a semi-human one. A text interface may make it easier to overlook the human/computer difference. Such a result has interesting implications – in situations in which a machine is being compared to a human, it may be that the medium through which the message is delivered will be significant. From the work discussed above, we make the following propositions:

P3: The more positive affect a machine can generate in a human, the more it will be trusted.

P4: The medium and valence of communication will influence both affect and initial trust.

4. Initial Trust

In the domain we are interested in, collision avoidance, it is often the case that pilots and air traffic controllers are introduced to each other for the first time right before they need to make a crucial decision. For this reason, we start by summarizing a model of initial trust developed by McKnight, Cummings and Chervany [31]. This model is based on trust of humans, not machines. We are seeking to parameterize this initial trust model according to the type of entity to be trusted.

This model itself blends several streams of research. We look at each stream in turn, and then discuss how the model can serve as a base for the model which is the subject of this paper.

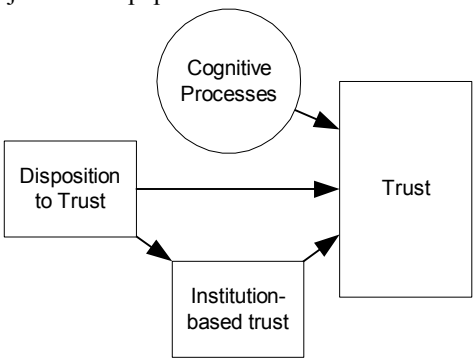

Figure 2. Simplified from McKnight et al. [31].

McKnight, Cummings, and Chervany propose that initial trust forms through one's disposition to trust, one's institution-based trust, and cognitive processes. Disposition to trust, the authors point out, encompasses a broad range of personality-based research, such as the work of Rotter [32]. Within disposition to trust, the authors differentiate between a general faith in humanity and a trusting stance. A trusting stance is much more calculated – one trusts even if one is not optimistic. We wonder if there is a corresponding general faith in machines – in other words, if those who trust humanity also trust machines. This leads us to the following research propositions:

P5: The disposition to trust machines can be measured.

P6: The disposition to trust machines is distinct from the disposition to trust humans; some individuals will be predisposed to trust one type of entity over another.

Within cognitive processes, McKnight, Cummings, and Chervany differentiate between categorization processes and illusions of control processes.

Categorization processes encompass stereotyping. The authors cite research which explores the impact of gender on trust [33]. We are interested in whether machines, and within that, types of machines, are also subject to stereotyping. We think it is likely that this is the case; brands of cars are often stereotyped on their reliability. This leads to the following proposition:

P7: Stereotyping cognitive processes will influence initial trust in a machine the same way it influences initial trust in a human, although the categories for stereotyping will be different.

Machines do not normally have gender or ethnicity as attributes. But model type, brand, and age are the kinds of attributes that might be subject to stereotyping. (When gender is added to a computer, people do stereotype the computer according to gender [34]).

The categorization processes of the initial trust model also include unit grouping, which is the tendency to trust others like oneself. The authors cite research on the tendency of humans to trust those in the same group [35]. On first glance, we think this process would seem to favor trust in people over machines. Yet it may be that certain machines may be perceived as members of a unit.

P8: Machines that are perceived as part of a unit group will be more highly trusted.

P9. Machines with a capability for autonomous behavior and affect generation will be more likely to be categorized as part of a unit group.

It may be that machines will always be considered outside the unit group, in which case the establishment of mixed teams of agents and humans will be impossible. Our proposition is more optimistic.

The categorization process includes reputation categorization. We think this is related to the problem of stereotyping. Reputation about humans is generally linked to information about a particular person. Machines are manufactured, and therefore we expect few individual differences from within a particular class of model. However, product models do acquire reputations, often based on anecdotal information.

In addition to categorization, McKnight, Cummings, and Chervany also discuss research on illusion of control processes [36]. People will often test other people conversationally – if they can provoke a desired response, then they often will trust the person. We are interested if this maps to machines. Machines in general are perceived

as being under our control, so there is no sense in testing them. Autonomous machines are different, however. We believe that testing with autonomous machines (for example asking the machine for feedback on progress) will lead to greater trust in the machine. Therefore we offer the following proposition.

P10: Autonomous machines which respond to control tests will be more trusted than autonomous machines that do not respond to such tests.

The third area of trust explored by McKnight, Cummings, and Chervany is that of institution-based trust. People will form trusting beliefs if the situation appears normal and if they believe that the organization provides structural safeguards, such as punishment of transgressions [37]. It makes no sense to punish a machine. But institution-based trust in a machine may increase if a human is associated with the machine. For example, if a pilot knows that the programmer of the latest revision of his control system will be punished if that programmer is negligent, then trust may increase.

P11: Machines that have identifiable humans accountable for their performance will be trusted more than machines which have no identifiable person accountable.

5. Individual Differences

Individual differences are a key driver of performance across many domains. We are particularly interested in the role of personality of individuals acting as part of a team in which some of the members are humans and some of the members are machines. The Five Factor Model of personality posits the following personality traits as a general explanatory framework for interpersonal behavior [38].

Openness has been associated with being imaginative, cultured, curious, original, broad minded, intelligent, and artistically sensitive. *Stability* is the extent to which an individual is calm, enthusiastic, poised, and secure. *Agreeableness* is the extent to which team members are good-natured, gentle, cooperative, forgiving and hopeful. *Conscientiousness* is the extent to which an individual is careful, thorough, responsible, organized, and planful, as well as hardworking, achievement oriented, and persevering. *Extraversion* is associated with being sociable, talkative, assertive, and active.

We are especially interested in Openness, Agreeableness and Conscientiousness as individual differences that may play a role in the behavior of subjects interacting with a machine-human team. Conscientiousness, for example, has been associated with the tendency to be cautious and avoid mistakes. It may be that highly conscientious individuals will be more likely to follow instructions under conditions of uncertainty and risk than individuals who are low on the same trait [39, 40]. Openness has been studied in the context of new product development teams [41, 42] and decision making teams. In the latter study, the investigators found that openness moderated the effectiveness of computer assisted decision making. More open individuals made better decisions under conditions of computer mediated communication. Agreeableness is a construct that has been directly related to trusting others [43] but has not been studied with respect to trust in machines. Agreeableness has also been found to be associated with effective team performance [44-48] but has not been generalized to teams in which one or more members are machines.

P12: People will exhibit reliable individual differences in their trust of machines

P13: High degrees of openness and agreeableness will predict a high predisposition to trust machines.

6. Training and Stress

The influence of individual differences, autonomy and cognitive processes on trust and subsequent decision making can be affected by two other factors, training and stress. In many cases we train in order to counteract the effects of stress [49, 50]. There is also evidence that providing people with appropriate preparatory information prior to a stressful event can reduce negative performance due to stress [51]. A large literature exists on how stress changes decision making, and in particular the decision making of pilots [52-57]. As a generalization, decision making is worse under conditions of time pressure. Some studies show that the choice to use automation is influenced by our own self-confidence in a situation – the less self-confidence, the more likely we are to choose automation. The more stress there is, the worse decision making becomes, and the more likely that individual differences and cognitive processes such as

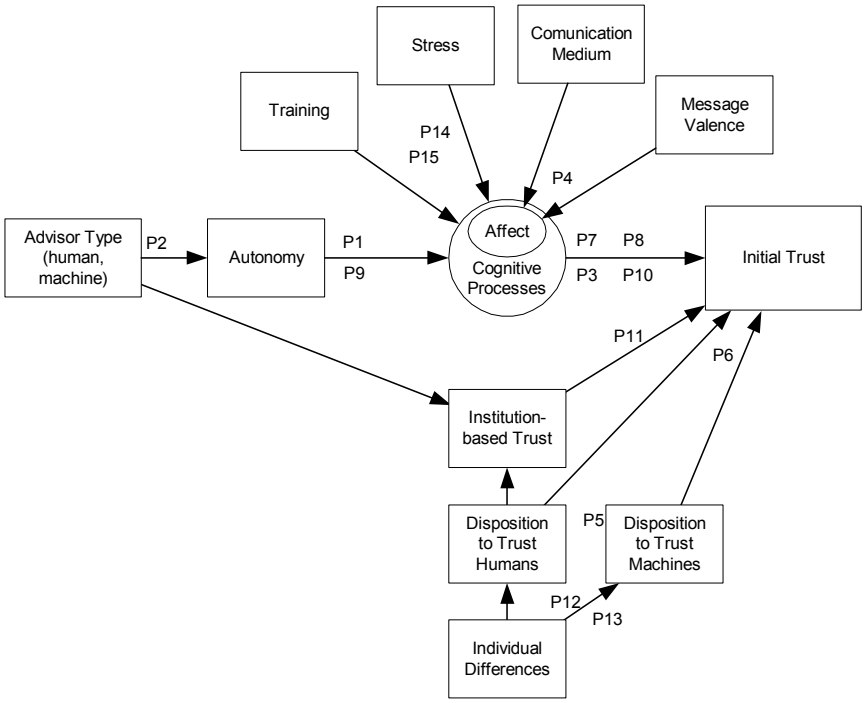

Figure 3. The proposed model.

stereotyping will surface, which may lead to an inappropriate level of trust.

P14: Under stress, both cognitive processes and individual differences will exert a stronger effect on initial trust than when stress is low or absent.

P15: The direct effects of stress on initial trust in machines will differ for trained and untrained individuals.

7. The Proposed Model

Synthesizing the theories we have discussed, we can posit a model of initial trust that is parameterized by the source of advice, machine or human. Along the bottom of figure 3 are the stable constructs related to personality.

Along the top are constructs which influence cognitive processes. These may enforce or negate each other – for example, we expect high valence advice, delivered through shouting, to accentuate the effects of stress, while training will dampen the effects of stress.

An overall situation in which a human and a computer provide conflicting advice can be represented as shown in figure 4.

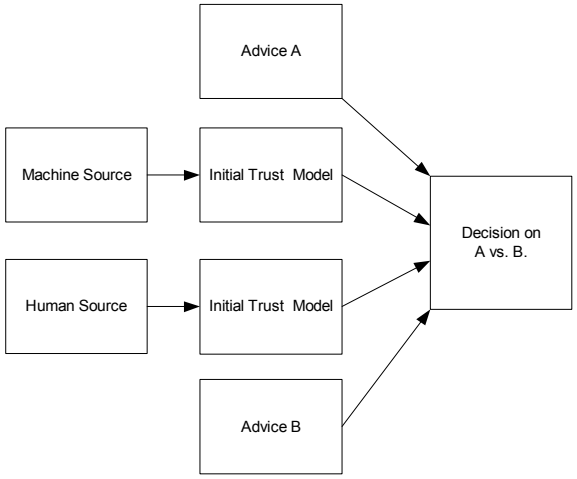

Figure 4. Use of the model

The initial trust model of figure 4 has as an input the source of advice, machine or human. The model outputs a value in each case. These values are used in making a decision between advice A, produced by a machine, and advice B, produced by a human.

In order to further explain the issues that model can be applied in studying, we present two scenarios, one concrete and historical, the other anticipatory.

8. An Accident Analysis

This scenario is based on the plane crash over Switzerland, July 1, 2002. Our information is drawn from news coverage [58, 59], and an accident description report from the Aviation Safety Network [60].

Two planes approach each other at right angles at the same altitude. Both planes ran a Traffic Alert and Collision Avoidance System, (TCAS) which operated properly. One plane is flown by a European pilot, the other by a Russian pilot. Both pilots are instructed by ground control to do the opposite of what the TCAS system tells them to do. The European pilot obeys the TCAS system. The Russian pilot obeys the ground controller. They crash.

We illustrate the scenario in figure 5. We label the TCAS transponders as A1, A2, the pilots as P1, P2, and the air traffic controller as C.

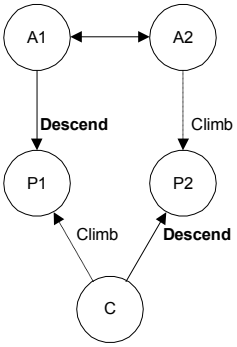

Figure 5. Contradictory instructions.

It is not clear from news reports if the controller ever issued the instruction to the European pilot, P1, to climb. If he didn't, then the European pilot's decision was easy – without a specific alternative, he would have obeyed TCAS.

We know more about the interaction between the Russian pilot and ground control. The pilot was faced with a quandary – TCAS told him to climb, the ground controller told him to descend.

Once the pilot found himself in the situation in which those two recommendations were in conflict, he now was faced with four possible outcomes, depending on both his decision and the decision of the other pilot, as shown in figure 6. In two situations, either TCAS or the ground controller is obeyed by both pilots. In two other situations, each pilot obeyed a different master.

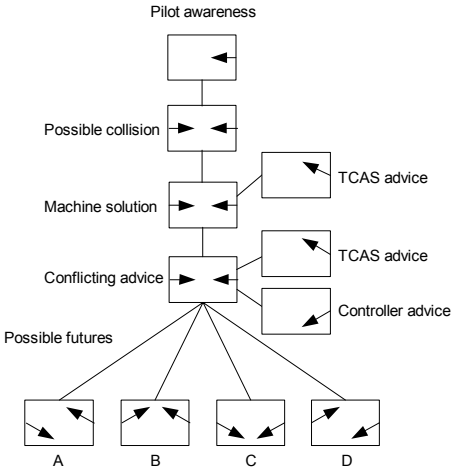

Figure 6. The possible situations

The Russian pilot would probably assume ground control had talked to the other pilot, and the other pilot was also going to do the opposite of what TCAS said. But the Russian pilot couldn't be sure – it was possible that ground control had not talked to the pilot, that the other pilot might disregard ground control instructions, or even that ground control had confused the two planes.

News reports focused on the idea that Europeans are more strenuously trained to always favor TCAS instruction, while the Russian policy gives more room to listen to ground control advice. In figure 7 we show the predispositions that may have influenced the decision making.

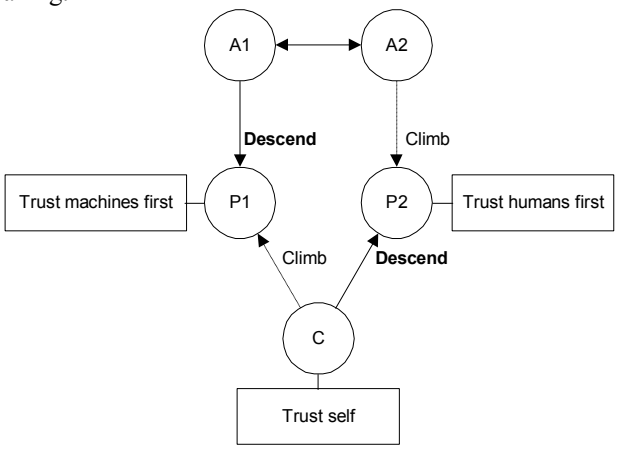

Figure 7. Representing simple heuristics

The implications of the news reporting are that if training for pilots was consistent world-wide, the accident wouldn't have happened. We are not so sure. It could be that in the above situation, a certain percentage of pilots will always decide in favor of the human, not the

machine. In other words, from our model, cultural differences or individual differences may create a disposition to trust humans over machines.

In other work on automation, accidents are sometimes traced to differences in models – a pilot may think the machine will behave in a different way than it does [61-63]. In this situation, the problem is probably not with the pilots' model of TCAS. Rather, it has to do with each pilot's model of what the other one will do.

9. UAV Collision Avoidance

The accident above was seen as an anomaly – but we think that such confusions have the potential to become more frequent as flight automation increases. We look now at a situation in which a human pilot is on a collision course with an unmanned aerial vehicle.

Such a situation is probably going to alarm both the human pilot and the air traffic controller. For the automated vehicle doesn't have the same survival motivations as the pilot does. Communication with the vehicle is not going to be the same as communication with another pilot. We show in figure 8 a set of possible communication links between a set of humans and agents involved in avoiding a potential collision.

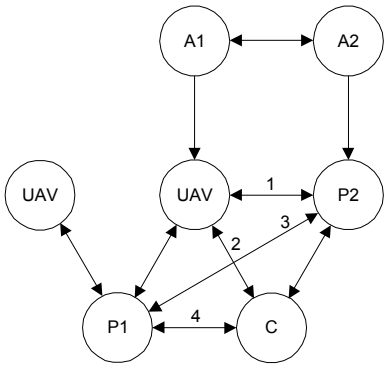

Figure 8. UAV communication

The pilot may resist talking to a UAV, along link 1, although this is the most direct interface. This resistance may be sensible, for the UAV will not be capable of human-level speech recognition, and the pilot may not have the time or inclination to interact using a different kind of interface.

The aircraft controller may also resist communicating to the UAVs along link 2 – it is unlikely the UAVs will be able to respond to arbitrary verbal commands, so some specialized interface – and training – would be needed by the air traffic controller. The pilot may prefer to talk along link 3 to the remote pilot of the UAVs. And the controller may also prefer to talk to the remote pilot along link 4. But the remote pilot will be controlling multiple UAVs, and will want the pilot in the manned plane to just let TCAS, labeled as A1 and A2, negotiate any changes in flight path.

It needs to be mentioned that the remote pilot probably will not have the same level of flight training as a pilot – and that, in some contexts, the pilots may not want to talk to the remote pilot. From the described model of initial trust, those outside the *unit group* are trusted less than those inside. Pilots may view talking to remote pilots as preferable to talking to machines – that even though they may be perceived as outside the Unit Group, humans are perceived as closer to that group than machines.

Generalizing this discussion, people prefer to talk with people rather than run specialized interfaces, due to aspects of initial trust of machines versus humans. Yet from an overall perspective, the less people need to talk to the remote pilot in this system, the better, as any overloading of the remote pilot increases the possibility of error. From our model, we hypothesize that the communication medium may have an effect. It may be that text-based communication, which will hide the differences between machine and human, may be the preferable way of lessening the tendency to trust the remote pilot over the machine.

The crux of the problem is the wide gap between a human being and an autonomous machine. Lawrence and Lorsch [64] showed that in conditions of great difference, integration happens through diplomatic go-betweens who have characteristics that are halfway between the un-integrated groups. In applying this theory, we might observe that what is needed is something halfway in between the pilot and the UAVs. The remote pilot currently serves such a function. This suggests that the natural tendency will be to deluge the remote pilot with requests. And that the design of doctrine, training, software and overall architecture would need to actively seek ways of avoiding such a deluge. The research we propose will take a step in the direction of understanding how as a society we should consider handing these problems before they become prevalent.

10. Experiment development

We plan to do a series of experiments in which a subject or, in some cases, more than one subject, will be placed in the role of either a pilot or an air traffic controller. We are interested in two dependent variables.

First, we expect our independent variables to influence the level of trust that subjects have toward machines and we expect trust to be strongly related to the decision (machine recommended vs. human recommended) made by subjects.

The measurement of trust is in itself a research area. Muir proposed a two-dimensional approach [65]; Lee and Moray argue the dimensions are not independent [66]. Marsh presents a formal measure involving risk [67]. McAllister defined trust as, "the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another" [68]. McAllister suggested that interpersonal trust can be categorized into two different dimensions: cognitive and affective. *Cognitive* forms of trust reflect issues such as the reliability, integrity, honesty, and fairness of a referent. McAllister showed that cognitive trust in peers was associated with the reliable performance of the peer and the extent of interaction with the peer. Status based on formal credentials, organizational role, etc. had a weak relationship with trust. *Affective* forms of trust reflect a special relationship with the referent that may cause the referent to demonstrate concern about one's welfare. The level of interaction with a peer had a major influence on the level of affective trust. It may be that machines which show affect are trusted along this dimension also.

Our research will include both experimentally manipulated variables and individual differences. We plan to measure each subject's personality using a measure of the five-factor model of personality [39] and, in addition, assess each subject's predisposition to trust a machine versus a human. We plan to experimentally manipulate a number of other independent variables. For example, we plan to vary the extent to which the machine has human characteristics and then look at the multiple regression of decision-making on the two types of trust, affective and cognitive. Our hypothesis is that as machines take on more human characteristics the regression weight for affective trust will increase and the regression equation will become more similar to a regression equation for a human information source.

Following the ideas of Parasuraman, Sheridan and Wickens [13] the degree of autonomy of the machine agent could be manipulated by allowing the subject varying degrees of control. The level of autonomy of a machine can be manipulated and the effects on trust and decisions can be observed. Of interest is the form of the relationship between autonomy and trust. For example, is trust a linear function of autonomy? Or is trust related to autonomy in some more complex way?

We plan to manipulate various characteristics of the machine (e.g., speech) to determine whether these characteristics elicit more positive or negative affect. We will test the proposition that affect is related to trust in the machine and subsequent decisions.

The extent to which initial trust in machines determines the decision made by subjects under various conditions will be explored in our experiments.

The relationships between personality and initial trust in machines will be determined. Other individual differences (e.g., experience with computers) will also be correlated with initial trust.

The role of training on initial trust and subsequent decisions will be explored in combination with other variables. For example, to what extent does training mitigate the influence of individual differences on trust in machines? To what extent does training mitigate the effects of stress on trust and decision making?

11. Conclusion

We have described an approach to investigating the effects of machine autonomy on human behavior. Our interest is in people's reactions to autonomous machines, rather than the nature of the machines themselves. We think that as a society we understand very little about how things might change as machines become more autonomous. As individuals, we may be reluctant to delegate, or we may over-delegate. We may cede decisions to the machine when we should intervene, and we may intervene unsafely when the machine should be left alone.

The issues we raise are large, and the related literature spans many different disciplines. We have synthesized this literature. More importantly, we have focused on a domain, collision avoidance, which lends itself to human experiments. We have shown how the broad questions we have raised can be specifically tested.

We have raised important and testable research questions. We have surveyed and synthesized a wide range of approaches to these questions. And we have outlined a clear, systematic way of beginning to answer these questions. These answers will be useful to a broad range of researchers.

The problem of how we delegate and intervene with a machine affects society as a whole. Our culture is built on our ability to trust and coordinate with each other. The introduction of autonomous machines means we have to look harder at trust, delegation, and intervention. Without thoughtful research, we will find ourselves reacting to changes brought about by technology. We will be better off if we anticipate – then we will be prepared to design our machines, as well as the training and policies which surround the ways we interact.

12. References

[1] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International Journal of Human-Computer Studies*, vol. 40, pp. 153-184, 1994.

[2] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, pp. 697-718, 2003.

[3] M. Lewis, "Anticipation, delegation, and demonstration: Why talking to agents is hard," in *Cooperative Information Agents III*, vol. 1652, *Lecture Notes in Artificial Intelligence*, M. Klusch, O. Shehory, and G. Weiss, Eds, Springer-Verlag, 1999, pp. 365-389.

[4] T. Lenox, L. Roberts, and M. Lewis, "Human-Agent interaction in a target identification task," IEEE International Conference on Systems, Man, and Cybernetics, 1997.

[5] K. Sycara, M. Lewis, T. Lenox, and L. Roberts, "Calibrating trust to integrate intelligent agents into human teams," 31st Annual Hawaii International Conference on System Sciences, 1998.

[6] K. S. Barber and C. E. Martin, "Agent Autonomy: Specification, Measurement, and Dynamic Adjustment," Autonomy Control Software Workshop, Seattle, Washington, 1999.

[7] R. Falcone and C. Castelfranchi, "The human in the loop of a delegated agent: The theory of adjustable social autonomy," *IEEE Transactions on Systems, Man, and Cybernetics - Part A,* vol. 31, pp. 406-418, 2001.

[8] L. Foner, "What's an Agent, Anyway?" MIT Media Lab 1997.

[9] T. B. Sheridan and W. R. Ferrell, Man-machine systems; information, control, and decision models of human performance. Cambridge, Mass.: MIT Press, 1974.

[10] T. B. Sheridan, *Telerobotics, automation, and human supervisory control*. Cambridge, Mass.: MIT Press, 1992.

[11] T. B. Sheridan, "Some Musings on Four Ways Humans Couple: Implications for Systems Design," *IEEE Transactions on Systems, Man, and Cybernetics - Part A,* vol. 32, pp. 5-10, 2002.

[12] C. E. Billings, *Aviation automation: the search for a human centered approach*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers, 1997.

[13] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A,* vol. 30, pp. 286-297, 2000.

[14] H. Hexmoor, "Case Studies of Autonomy," Thirteenth International Florida Artificial Intelligence Research Symposium Conference, Orlando, Florida, 2000.

[15] T. Inagaki, "Situation-adaptive degree of automation for system safety," 2nd IEEE International Workshop on Robot and Human Communication, Tokyo, Japan, 1993.

[16] S. Lewandowsky, M. Mundy, and G. P. Tan, "The Dynamics of Trust: Comparing Humans to Automation," *Journal of Experimental Psychology: Applied*, vol. 6, pp. 104-123, 2000.

[17] M. Williams, "In whom we trust: Group membership as an affective context for trust development," *Academy of Management Review*, vol. 26, pp. 377-396, 2001.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, (in press).

[19] C. Nass and J. Steuer, "Voices, Boxes and sources of Messages: Computers as Social Actors," *Human Communication Research*, vol. 19, 1993.

[20] C. Nass, J. Steuer, L. Henriksen, and D. C. Dryer, "Machines, Social Attributions and Ethopoeia: Performance Assessments of Computers Subsequent to 'Self' or 'Other' Evaluations," *International Journal of Human-Computer Studies*, vol. 40, pp. 543-559, 1994.

[21] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer, "Can Computer Personalities be Human Personalities?" *International Journal of Human-Computer Studies*, vol. 43, pp. 223-239, 1995.

[22] T. Bickmore, "Relational Agents: Effecting Change through Human –Computer Relationships," in *Media Arts and Science*: MIT, 2003.

[23] B. Friedman, "It's the computer's fault: reasoning about computers as moral agents," CHI 95, Denver, Colorado, 1995.

[24] B. Friedman and H. Nissenbaum, "Software Agents and User Autonomy," 1st international conference on Autonomous agents, Marina del Rey, California, 1997.

[25] C. Breazeal, "Robot in Society: Friend or Appliance?" Agents99 Workshop on Emotion-Based Agent Architectures, Seattle, WA, 1999.

[26] B. Friedman, P. H. Khan, and D. C. Howe, "Trust online," *CACM*, vol. 43(12), pp. 34-40, 2000.

[27] H. Prendinger and M. Ishizuka, "Agents That Talk Back (Sometimes): Filter Programs for Affective Communication," 2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction, in conjunction with User Modeling 2001, Sonthofen, Germany, 2001.

[28] C. F. DiSalvo, Gemperle, F, Forlizzi, J, & Kiesler, S, "All robots are not equal: The design and perception of humanoid robot heads," Designing Interactive Systems, Dis2002, London, 2002.

[29] C. Nass and K. M. Lee, "Does computer synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction," *Journal of Experimental Psychology: Applied*, vol. 3, pp. 171-181, 2001.

[30] S. Kiesler, S. S, L, and K. A. Waters, "A Prisoner's Dilemma Experiment on Cooperation With People and Human-Like Computers," *Journal of Personality and Social Psychology*, vol. 70, 1996.

[31] D. H. McKnight, L. L. Cummings, and N. L. Chervany, "Initial Trust Formation in New Organizational Relationships," *Academy of Management Review*, vol. 23, pp. 473-490, 1998.

[32] J. B. Rotter, "A new scale for the measurement of interpersonal trust," *Journal of Personality*, vol. 35, pp. 651-665, 1967.

[33] J. Orbell, R. Dawes, and P. Schartz-Shea, "Trust, social categories, and individuals: the case of gender," *Motivation and Emotion*, vol. 18, pp. 109-128, 1994.

[34] E.-J. Lee, "Effects of "gender" of the computer on informational social influence: the moderating role of task type," *International Journal of Human-Computer Studies*, vol. 58, pp. 347-362, 2003.

[35] R. M. Kramer, M. B. Brewer, and B. A. Hanna, "Collective Trust and Collective Action: The decision to trust as a

social decision," in *Trust in Organizations: Frontiers of Theory and Research*, R. M. Kramer and T. R. Tyler, Eds. Thousand Oaks, CA: Sage, 1996, pp. 357-389.
[36] E. J. Langer, "The Illusion of Control," *Journal of Personality and Social Psychology*, vol. 32, pp. 311-328, 1975.
[37] L. G. Zucker, "Production of trust: Institutional sources of economic structure," in *Research in organizational behavior*, vol. 8, B. M. Straw and L. L. Cummings, Eds. Greenwich, CT: JAI Press, 1986, pp. 53-111.
[38] M. R. Barrick and M. K. Mount, "The Big Five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, vol. 44, pp. 1-26, 1991.
[39] P. T. Costa and R. R. McCrae, "Revised NEO Personality Inventory manual," Psychological Assessment Resources, Odessa, FL 1992.
[40] L. R. Goldberg, "A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models," Personality psychology in Europe, Tilberg, the Netherlands, 1999.
[41] R. Reilly, G. Lynn, and Z. Aronson, "The role of personality in new product development team performance," *Journal of Engineering and Technology Management*, vol. 19, pp. 39-58, 2002.
[42] R. Reilly, Z. Aronson, and G. Lynn, "New Product Development Team Performance: The role of team member personality," Academy of Management Annual Convention, Seattle, 2003.
[43] K. DeNeve and H. Cooper, "The Happy Personality: A Meta-Analysis of 137 Personality Traits and Subjective Well-Being," *Psychological Bulletin*, vol. 124, pp. 197-229, 1998.
[44] L. M. Hough, "The "Big Five" personality variables-construct confusion: Description versus prediction," *Human Performance*, vol. 5, pp. 139-155, 1992.
[45] S. L. Kichuk and W. H. Wiesner, "The Big Five personality factors and team performance: Implications for selecting successful product design teams," *Journal of Engineering and Technology Management*, vol. 14, pp. 195-221, 1997.
[46] M. R. Barrick, G. L. Stewart, M. J. Neubert, and M. K. Mount, "Relating member ability and personality to work-team processes and team effectiveness," *Journal of Applied Psychology*, vol. 83, pp. 377-391, 1998.
[47] M. J. Stevens, R. G. Jones, D. L. Fischer, and T. D. Kane, "Team performance and individual effectiveness: Personality and team context," 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, 1999.
[48] G. A. Neuman and J. Wright, "Team effectiveness: Beyond skills and cognitive ability," *Journal of Applied Psychology*, vol. 84, pp. 376-389, 1999.
[49] C. E. Hartel and G. Hartel, " SHAPE-Assisted Intuitive Decision Making and Problem Solving: Information-Processing-Based Training for Conditions of Cognitive Busyness," *Group Dynamics: Theory, Research, and Practice*, vol. 1, pp. 187-199, 1997.
[50] J. H. Johnston, J. E. Driskell, and E. Salas, "Vigilant and Hypervigilant Decision Making," *Journal of Applied Psychology*, vol. 82, pp. 614-622, 1997.
[51] C. M. Inzana, J. E. Driskell, E. Salas, and J. H. Johnston, "The effects of preparatory information on enhancing performance under stress," *Journal of Applied Psychology*, vol. 81, pp. 429-435, 1996.
[52] R. J. Adams, "How expert pilots think: Cognitive processes in expert decision making," DOT/FAA/RD-93/9. Washington, DC. (NTIS No. AD-A265 356/61NZ). 1993.
[53] L. H. Franklin, E, "An emergency situation simulator for examining time pressured decision making," *Behavior Research Methods, Instruments and Computers*, vol. 25, pp. 143-147, 1993.
[54] L. Reder and R. L. Klatzky, "The Effect of Context on Training: Is Learning Situated?" Carnegie Mellon CS-94-187, 1994.
[55] M. R. Endsley, "Toward a theory of situational awareness in dynamic systems," *Human Factors*, vol. 37, pp. 32-64, 1995.
[56] G. Klein, "The current status of the naturalistic decision making framework. I," in *Decision making under stress: Emerging themes and applications*, R. Flin, Ed. Brookfield, VT: Ashgate, 1997.
[57] P. A. Craig, "Improving pilot decision making in situations of high stakes, high stress and time pressure," Tennessee State University, 1998.
[58] BBC, "What Might Have Gone Wrong?" http://news.bbc.co.uk/1/hi/world/europe/2082331.stm 2002.
[59] BBC, "Crash Fuels Europe Airspace Debate," http://news.bbc.co.uk/1/hi/world/europe/2089030.stm 2002b.
[60] AviationSafetyNetwork, "Accident Description July 1st 2002," http://aviation-safety.net/database/2002/020701-0.htm 2002.
[61] J. Rushby, "Using Model Checking to Help Discover Mode Confusions and Other Automation Surprises," *Reliability Engineering and System Safety*, vol. 75, pp. 167-177, 2002.
[62] J. C. Campos and M. D. Harrison, "From Interactors to SMV: A Case Study in the Automated Analysis of Interactive Systems," Department of Computer Science, University of York. Technical Report YCS-99-317, 1999.
[63] C. S. Oliveras, "Systems, Advisory Systems and Safety," Department of Computing Science, University of Newcastle CS-TR: 774, 2002.
[64] P. R. Lawrence and J. W. Lorsch, *Organization and environment; managing differentiation and integration*. Boston: Harvard University, 1967.
[65] B. Muir, "Trust In Automation: Part 1. Theoretical Issues in the Study of Trust and Human Intervention in a Process Control Simulation," *Ergonomics*, vol. 39, pp. 429-460, 1994.
[66] J. D. Lee and N. Moray, "Trust and allocation of function in human-machine systems," *Ergonomics*, vol. 35, pp. 1243-1270, 1992.
[67] S. Marsh, "Formalising Trust as a Computational Concept," in *Department of Mathematics and Computer Science*: University of Stirling, 1994.
[68] D. J. McAllister, "Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of Management Journal*, vol. 38, pp. 24-59, 1995.