

CS4IB3

A Survey of Human Interaction with AI Systems

Dan Zingaro

In a field so diverse and encompassing as Artificial Intelligence, it should not be surprising that people may have strikingly different views on the role that AI plays in their lives. Psychologists, computer scientists, philosophers, and the general public, all have their own interpretations of what the field means to them. Some may believe that “true” AI has not arrived yet, and only has its place in sci-fi TV shows. Others may equate AI with virtual reality systems, or home video games, where AI is given practical applicability. However, as will be demonstrated in this paper, AI is already having a much larger effect on people than many may realize. And, as the field progresses further, AI will be making its way out of professor’s stuffy offices and move even more directly into our homes and work places.

This paper aims to introduce the reader to several aspects of the phenomenon of humans interacting with Artificial Intelligence systems. A natural first question to ask is, “why are we using AI systems at all?” Can they really be useful for anything? This question will be tackled first, as the numerous uses of AI as applied to a person’s lifestyle, are outlined. The topic of data mining is also briefly outlined, because it is this technique which gives AI systems much of their power and flexibility. Then, the psychological aspect of trust (or belief) in these so-called Expert Systems will be assessed. Do people feel comfortable leaving decisions to these systems? How much faith do people have that these decisions are correct? Third, an attempt will be made to illustrate techniques for implementing AI systems that interact with humans, in a safe and secure way. To go to the extreme, people will obviously not let a robot run around their homes if they didn’t believe that it was extremely safe to do so. Furthermore, people must generally feel that the infusion of AI systems is a positive experience—they must feel secure in their dealings with these systems. Finally, the rather influential idea of using AI systems as teachers will be surveyed. Can AI systems make good teachers? What are the techniques for ensuring that learner’s derive maximum benefit from their sessions with the teacher? Could this, one day, affect the traditional school system? At the end of this paper, it is the author’s hope that the reader will possess a new respect for the boundless potential of AI systems, when they are being used both correctly and incorrectly.

Part 1. Motivational Aspects

Let us now sample the motivational aspects of the field, briefly from an academic point of view, and then more extensively from a general-public point of view. The field of Data Mining has produced some very useful research, and is in fact the backbone of many AI systems. These days, it is not difficult to observe the wealth of data which exists electronically—a quick read of google.com shows that well over 8 billion pages are indexed by that search engine. However, for an intelligent system to operate on the data, sheer terabytes of accumulation are not enough. In fact, there must be a way of extracting the useful information from this huge database (Zhou, 2003). This technique is known as Data Mining, and, since the 1980’s, has been studied in great depth by various disciplines (Zhou, 2003). In his review paper, Zhou collects the views of data mining by those studying it from the database, learning, and statistical, perspectives. This is one example of the previously unsubstantiated claim in this paper that AI is being studied at a variety

of levels. Those interested in database theory are obviously interested in data mining, since it is essentially the utilization of large databases. Those interested in machine learning want to study techniques for machines learning (or extracting useful data) from these data collections. Finally, statisticians are interested in errors that the machines make from the contained data, extracting statistically meaningful data, and the like. Regardless of the perspective taken though, it is evident that data mining provides a powerful means for systems to acquire new knowledge. While some may call this “learning”, the taxonomy given is irrelevant—the important point is that, with the backbone of data mining, AI systems can simulate the process of learning in humans. Many examples of AI systems interacting with humans will now be given, and, because learning is a prerequisite of intelligence, they all rely heavily on the aforementioned techniques.

Telepresence: Being There so you Don't Have to Be

Let's begin with a foretelling study by Agah, Tanie, Ohkawa, and Iwamoto (1997). These experimenters sought out to show that it would be possible to control a remote robot, and have it explore a museum at a person's command. This enters into the field of telepresence, which refers to the human being in a different (remote) location than the robot. What is remarkable about this particular design, however, is that, along with standard methodologies for input, other interesting techniques were used. One was through a head-mounted display system on the user, which tracked head motion, and interpreted it as commands for the robot to carry out. For instance, as illustrated in the article, the user moving his head to the left or right signified to the robot that it should turn in place, because the user's intent was to look in a different direction. In fact, human intention-understanding is pivotal for the human-robot communication in this case. By realizing what the user implicitly wants, the user is at the luxury of thinking less about how to instruct the robot, and more on enjoying the experience. In other words, it abstracts away from the command-oriented state of mind that people normally have when dealing with computers, and allows the end-user to behave more naturally. Taking a step back and looking at this study at face value, it may not seem very useful, and certainly no motivation for AI-human interaction. But consider the broader messages. Imagine having an AI system in a dangerous area, being controlled by humans in safety. Or, more mundanely, but equally helpful, think about the possibility of a robot moving around in a shopping area, being controlled by someone who can no longer do these daily activities themselves. Certainly, one can argue that this is something to get excited over, and an excellent reason for AI to become more intertwined in people's lives. Of course, safety issues must be taken into consideration, and they will be as we proceed.

Mobility

Let's now look at how AI systems can potentially be the next step in assistive mobility devices for the visually impaired, or those in wheelchairs. In an ongoing, recent work by Mori, Kotani, and Kiyohiro, (1994), the current state of mobility aides was found unsatisfactory. To see why, realize that the ultimate goal of a mobility aide is to enhance the user's safety, and allow him to successfully reach his destination. It is now not difficult to see that current, standardized devices are lacking. For example, those who use

canes only can determine what is several meters ahead, which is not very useful in unfamiliar territory. Furthermore, guide dogs have historically been hard to train, and previous attempts at an AI solution have failed due to their employed method of feedback. These authors are in the process of developing a new robot-travel-aide (RTA) with several concepts, rooted in the field of AI. Most notably, it incorporates object-avoidance techniques, and uses a knowledge base to make many of its decisions. The RTA also includes a bar that the client can use to determine the regularity (or lack thereof) of the ground ahead. The authors make a point of stating that the RTA does not make all decisions on its own, but that this is not necessarily a bad thing. For example, sometimes it is impossible for a robot to know which of several alternatives to take, should the standard path be blocked. At these points, the authors reason, it is probably better to rely on human judgment, and let the robot take over again once a decision has been made. Putting this issue to the side for the time being, it is clear that a system such as this could be of great significance in giving people greater independence.

Virtually Real

Perhaps what comes to mind most often when thinking about artificial intelligence is the concept of virtual reality. As its name implies, virtual reality attempts to convince the user that the situation is real, when in fact it is only being produced virtually. Video games have tried to do this for many decades now, and so the entertainment value of AI systems is no longer a surprise to most people. These video games present an imaginary world, and allow the gamer to interact with it through a set of accepted commands. The realism has increased drastically in recent years, as these systems have moved from 8-bit graphics where one's imagination must work overtime to fill in the details, to systems which match, or eclipse, the realism seen in today's movies. However, the entertainment aspect can go even further than this, as is evident in a Virtual Baseball System developed by Komura, Kuroda, and Shinagawa, (2002). These authors point out that it is difficult for athletes to train on their own—for example, how is a baseball player expected to hit pitches with no one throwing them to him? There are, of course, baseball video games, but these hardly emulate the real thing to a satisfactory degree. The solution, they argue, is through virtual reality; so this paradigm, in fact, can be useful for other aspects outside of entertainment. The hitter uses a real bat, equipped with an infra-red reflector, and stands in front of a screen where virtual pitches are thrown. The batter swings the bat at the appropriate time, trying to connect with the pitch; the pitch “knows” it has been hit by sensors behind the screen responding to the motion of the bat. What makes this system even more useful is that the batter can swing at pitches that would be thrown by all-stars of the game. It accomplishes this by using video sequences of different pitcher's motions when they throw the ball. Obviously, similar techniques could be applied to other sports; in fact, a virtual reality tennis system has already been developed (Kawamura, Ida, Wada, and Wu, 1995). As not to repeat the concepts, suffice it to say that the main idea of this tennis system is to mimic the force of the tennis ball, when it hits the racket. This, again, is a separating factor between traditional entertainment systems, and these more advanced paradigms. These and similar developments will, one day, allow athletes to practice their skills when it would not otherwise be feasible (for instance, during bad weather conditions, during the off-season when skilled opponents are unavailable, and so on).

Method of Communication

Let us take stock of the situation up to this point. By now it is obvious that various Artificial Intelligence systems have been deemed feasible to coexist with the general public, and to enhance their living in a variety of ways. One issue not touched upon yet, however, is the most effective means of communication with these systems. Researchers have to make an attempt to break the mold of current input techniques (keyboard, mouse, etc.), because they do not represent a natural communication channel. For these systems to be as seamless as possible, humans must have the ability to interact with them much as they interact with other members of the species. But even with this analysis, something is missing. Humans do not communicate one-dimensionally with others. To the contrary, people use what may be called implicit cues to effectively relay what they are trying to say. For instance, the same sentence may be interpreted as being sarcastic, or genuine, depending on the intonation of the voice. Facial expressions also quietly convey information which cannot possibly be so succinctly delivered with speech. Clicking a mouse or using a keyboard is thus miles from the elegant, specialized, and highly efficient method of communication that, until now, only humans have understood. If AI systems could not reproduce such techniques, then perhaps they would have a more difficult time becoming mainstream and being accepted. Certainly, for AI to flourish, communication channels must be further opened.

Fortunately, work by Iwano, Sukegawa, Kasahara, and Shirai, (1995) reveals that this process has already begun. The authors focus on the feelings of the human speaker, as they relate to the communication. Their thinking is that, if AI systems could interpret the human emotions, they could better “guess” what the human is requesting. The authors incorporate such cues as the movement of the head, and facial expression, and integrate them into a holistic emotion. Perhaps more interesting is that the authors also consider and use prosody, which refers to rhythm, stress, and other vocal characteristics. To refute any argument against the importance of the characteristics of speech, consider the differences between a casually spoken command, and a time-critical instruction. When time is of the essence, speech is necessarily quicker (more phonemes spoken per second) and may be more heavily stressed at some points. This is of great use to other humans and, of course, it would be fantastic if AI systems could benefit from it as well. Continuing, it is also known that there are several emotions which are considered universally produceable and recognizable across all cultures, including, for example, anger, happiness, and amazement. An AI system, such as the one currently being described, could conceivably recognize these emotions as well, and remain applicable to all of humanity. Admittedly, though, emotions can sometimes be ambiguous, making the decision process even more daunting for an AI system. Presently, human emotions have been divided into five types, and thereby classified by the system. For instance, looking at just one category, the authors state that the category of “dislike” is assigned if the human is interacting in a negativistic way. With this as an adequate foundation, let us leave the details of the study and look at the big picture. Why should the interpretation of human feeling motivate the general public to accept AI? Consider first, the potential of using facial expressions and other indications of feeling to use today’s current

technology. For those who are incapable of interacting with technology in the conventional way (for instance, due to a stroke causing loss of function), this may become a feasible method of interaction. Or, as another example, consider the effect it could have on our judicial system. Simple polygraphs could be enhanced by AI systems ready to read the person's feelings, and convey to the court how truthful he is being. Regardless of its use, the take-home message should be clear: these types of interpretations bridge the gap between humans and AI systems, and is another indication that AI should soon deserve a place in human society.

All Rise: The Court of Law

It may be instructive, now, to consider the potential fusion of the Court of Law and expert systems. From the data amassed so far, it is known that AI systems are capable of interpreting human actions, making what look to be "wise" decisions, and otherwise emulating a significant portion of humanity. Shouldn't there be a place in our courts and government organizations for such respectable technology? For instance, consider the typical court proceedings. A witness may be questioned several times over a period of days or weeks, and lawyers brood over the data in an all-out attempt to find inconsistencies. However, put another way, the lawyers are essentially presented with a mass of data (a database) and have to extract inconsistent bits of data. Is this not the basis for data mining? Unfortunately, work by Oskamp and Lauritsen, (2002) dims the lights on such possibilities. The authors state that no full-fledged AI system has been used in a real court of law, as of yet. However, lowering the ideal standards somewhat, the situation doesn't look so grim. Consider, for instance, the Australian company described in the article at hand: Softlaw. This company develops expert systems used by various sectors of the government. The department of Veterans' Affairs uses a compensation claims processing expert system. As well, the Department of Defense uses an expert system to make decisions about worker's compensation; for example, the amount of money received every week. So, contrary to the initial pessimism, AI systems are at least making a splash in governmental operations. That said, why should the general public be interested in this? The authors point to the work by Softlaw, and state that it allowed the government to process more claims, with greater efficiency, than was previously possible. This would undoubtedly benefit those waiting for government response on obtaining disability funding, worker's compensation cheques, and the like. For another example, let us turn, for the moment, to complete surmise. One issue which is constantly under public scrutiny is personal security. Websites wave online banners stating "certified against 99.9% of hacker crime", and the like, in order to ensure the viewers that purchases will be safe. Microsoft's Windows XP has recently incorporated a firewall and other security features so that people feel more confident in their computing tasks. Still, many people continue to stay clear of giving out their personal information online, because of accounts of fraud which permeate the media. Surely, any agent which will help in making people feel more secure, and ultimately increasing security, would be warmly received by all. Couldn't similar expert systems to the ones just be described be used by governments and other organizations to enhance security? When dealing with worker's compensation claims, couldn't the system be built to detect attempts at fraud? Hopefully, the not-so-distant future will illuminate these possibilities.

Robot-Assisted Surgery

It is now time to turn to an issue which may represent a moral dilemma for many people. Specifically, let us take a look at the usage of robots in the realm of surgery and other medical procedures. To appreciate the severity, consider the series of events which may transpire after an unsuccessful operation, resulting in injury or death. The family members, or friends, of the afflicted will certainly begin to question what happened, perhaps blaming the surgeon, or themselves, for the result. Now, consider the replacement of the surgeon with a robot, and assume the same, unfortunate results. Is this an improvement? How will everyone feel now? One thing that is clear is that the blame will shift from the surgeon, to a robot—but is this a good thing? Wouldn't people feel more comfortable blaming another of their own species than an abstract device made of metal and wires, known as a robot? Or, even worse, consider the possibility of allowing a family to choose between robot or human for the operation. Regardless of the choice, should something unexpected happen, the self-blame and brooding could be made even worse. Choose a human, with feelings, desires, a heartbeat, and someone with whom one can associate? Or choose a robot, hoping that this “advanced species” can do better than our own. This type of questioning can go on indefinitely, and so it is not clear that the introduction of robots into the hospital would be viewed positively by the general public. To break out of this circle, let us consider the hard facts, and see where that takes us.

These facts can be found in a study by Theodossy and Bamber (2003). These authors were interested in the type of surgery dealing with fixing deformities in the jaw and surrounding areas, called Orthognathic surgery. The study consisted of comparing standard, manual surgery, with surgery performed via the aide of a robot arm. The authors are careful to point out that this robot arm is passive, which means that it performs actions based on input from the surgeon. This is in contrast to an active robot, which does more of the decision-making, and is monitored, not ordered, by the surgeon. The results showed that the robot-assisted surgery was more effective and accurate in two out of three important areas (anteroposterior and vertical planes). The authors then rightfully conclude that, with the help of the robot arm, the planning and research portions of the surgery were improved. This last point deserves to be dwelled on for a moment. Note how the robot arm can be helpful in the initial, planning stages, with no reference to the system actually being used for the surgery itself. Surely, noone would fight against robots who could assist in the vital research stages of a pending surgery. In the least, this would allow the human to be better prepared to do the surgery alone, using the information gleaned from the robot arm. So, viewed in this way, the emergence of AI into the hospital can be seen as positive, and a step forward in medical technology! What is important is that AI can be of benefit in other ways than the “hands-on” portions of operations.

Before proceeding, one other general message can be obtained from this article. Everyone knows that humans make mistakes (in fact, many argue that this is the essence of humanity itself). As well, it is known that, barring various types of failures, machines only have the capacity to follow human instructions exactly as they were given.

However, going back to the hospital realm one last time, the article also states that the robot arm is more spatially accurate, and can achieve greater precision, than an adequately-equipped human. This is critically important, and should make it clear that whenever issues of timing, or accuracy, or fine detail, or “a steady hand” are necessary, the arena of AI should at least be given a look!

Robot Autonomy

Let’s return to the distinction between a passive robot and an active one. Relying on the intuitive meaning of active, a so-called active robot may be likened to an autonomous one. Autonomy is certainly important in the daily life of humans. From the realm of social psychology, it is known that the ability to make decisions and guide one’s own life has positive influences on a person’s happiness and health (Langer & Rodin, 1976). Obviously, robots do not require such human dependencies, but autonomy has much wider breadth than this. In short, an autonomous robot, who can do things on its own, may be more useful in certain situations (for example, a security robot protecting someone’s home would be quite useless if it required a confirmation before calling for help). What does the literature have to say on this?

Consider a study by Kobayashi, Une, Takashi, Mikurita, Ohya, and Yanagawa, (1994). These authors concerned themselves with an autonomous robot, which makes decisions while taking pictures. For instance, given a target to track, it makes decisions which optimize the quality of the pictures taken. The robot can decide when to move, which angle of rotation to use, the angle of elevation, and the amount of zooming. At face value, this development would be useful, as the authors state, for viewing a remote area, from a different place. This would encompass such areas as home security, unobtrusively taking videos of wildlife, and so on. But there is a much deeper meaning to the results of the study. The fact is, that robots can have some degree of independence—some aura of being autonomous.

The Right Decision

Now that this is established, why is this beneficial? Put differently, why should people care that robots can make decisions from within themselves? For part of the answer, let’s consult an extensive article written by Dautenhahn, (1995). The work, as the author notes, is highly motivated by the Social Intelligence hypothesis, so it is critical to understand this before continuing. This hypothesis attempts to find an empirically-accepted claim for how intelligence evolved. For instance, intelligence could have evolved in order to allow a species to survive (for instance, catch prey). Other possibilities exist, of course, but this theory claims that intelligence, at least in primates, grew out of the necessity to deal effectively with social situations (Whitin, 2000). As well, intelligence is built up around the idea that sociality is of critical importance. So if acting appropriately in social situations is the predecessor, and core, of modern intelligence, isn’t this where robot intelligence should begin as well? Surely, the adaptive evolutionary process must have selected for social intelligence for a reason, so it must somehow be important. This argument is the crux of the paper being discussed. The

author admits that social intelligence in robots may not be important, when other, more technical, solutions are available. But social intelligence is unavoidable in certain activities that robots may have to perform if made a part of our society. The author cites several examples, including the fact that robots may have to work in teams, or act in a socially respectable way if acting as an assistant to a human in a place of employment. Certainly, a socially defective robot will just not fit in as expected. There are, of course, other benefits that come with being socially intelligent. It is to these benefits that this paper will briefly survey. However, what may be more subtle is that, upon techniques being developed to give robots this skill, they will become even more useful in human society. It is for this reason that the subject at hand should concern the general public—because robots have the potential, upon being given the power to act socially appropriate, to put their technological prowess to work for us.

So what can a socially acceptable robot accomplish for humanity? Consider the example that the author gives of a service robot assisting a handicapped person. Certainly, gigabytes of knowledge on the precise notions embodying the disability would be beneficial. But could a robot really be of much help unless he had the ability to adapt to the individual person it was servicing? Further, consider the ubiquitous service robot—the robots which are supposedly going to run around our homes doing our dues. Humans are so used to forming relationships with other humans, so would it be rewarding for the human if it could not do the same with the robot? Probably not, as argued in this paper, because it is through the development of relationships that humans (and robots) can truly understand what someone else is feeling. In fact, the author makes a bold statement, declaring that today's robots act like socially deprived animals. (Such animals, being raised with no social influences, would accordingly have no idea how to act when presented with a social situation.) In this light, these systems, all of a sudden, seem as helpful as the family dog. Thus, it is clear that social intelligence is a prerequisite for robots to become true assistive devices.

Section Summary

At this point, our survey of the motivational aspects of integrating AI into human society, draws to a close. It is hoped that the mass of data collected will serve a dual purpose. On the one hand, the reader should realize that AI, as a discipline, has certainly evolved and matured substantially as a branch of computer science, psychology, and philosophy. On the other hand, it is also a lucrative research area, with immediate applicability to human society. Throughout the rest of the paper, it is assumed that there is consensus on these issues—mainly, that the reader accepts, or at least complies with, the current direction of AI integration.

In other words, let's assume that robots are here to stay, and it is our job to assess the potential risks involved. Throughout the first part of this paper, several references were made, pointing to these particular difficulties. Two important ones have been chosen and will be studied in depth.

Part 2: The Element of Trust

Let's begin with the concept of trust, or belief, between humans and AI systems. Is it important for humans to trust decisions made by these systems? What effects will trust have on the integration of AI into people's daily lives? What can implementers do to increase the level of trust felt by humans? To answer these questions, it will be instructive, first, to look at why trust is important in standard human populations. After all, if it is not essential for humans to trust other humans, surely it is no more important to trust robots.

The Evolution of Trust

Let us take an evolutionary perspective to show that, on the contrary, trust is mission-critical in the formation of relationships and friendships. Consider work by Barclay (2004). From previous work in the field, it is known that, if people have nothing to lose by doing so, then they will act in a selfish and non-altruistic manner to others. Consider the situation of finding money on the floor of a deserted hallway. Assuming no chance of being monitored, the probability of taking it, instead of reporting it, is high. This makes sense, because it is adaptive (and, unfortunately, evolutionarily correct) to do this. But what happens when reputation is thrown into the equation? In other words, does a person act differently, knowing that his actions will improve, or tarnish, the reputation he has in the community? In fact, in Barclay's work, it is evident that people may act in an altruistic fashion, for the sole purpose of being seen as trustworthy by others. People, in general, trust those who are altruistic, more than they trust those who do not act altruistically. Furthermore, a single person can not have working relationships with everyone else (for example, due to time and space constraints), so Barclay notes that humans may even be in a "competition to be most altruistic", in order to create these friendships. Diving slightly below the surface in the work, Barclay had groups of people play a public goods game, which involves contributing a chosen amount of money to a public good. By not contributing, and "free-riding", an individual may increase his personal wealth, but at the expense of the group. Later on, when these same participants played a "trust" game, those who were most altruistic in the original game were given the most money. The message is now clear. People like, trust, and prefer dealing with altruists. For some reason or other (reputation, in this work), people consciously work to be seen as trustworthy by others. Naturally, then, AI systems should work to be trusted by humans—to incur its inherent benefits, and, critically for our purposes, to be accepted, repaid, and treated as expected, by humans.

Trust in Traveler Information Systems

Let's look at one specific example of this phenomenon, as it directly applies to AI systems. One intuitive, and effective, way to have people lose trust in an AI system is for it to give back incorrect results. In fact, this holds on a variety of levels, and is not limited to the AI domain. A website giving information about TV listings would quickly lose

visitors if it continually listed incorrect information. News programs may lose clientele if they have a tendency to report incorrect or inaccurate information. So too, AI systems can lose their credibility if they make decisions which do not appear correct. In Fox (1996), the author attempted to assess just how important the trust factor between a human and an AI system is. While it may be argued that Fox's example is not a prototypical AI paradigm, this is inconsequential, as the general results are what is critically important. To facilitate the study, Fox focused on the Advanced Traveller Information System. These devices have become immensely popular in recent years, as they can provide drivers with route information, in order to arrive at the destination in the most appropriate manner. For example, they can help to avoid congestion, accidents, blocked routes, and the like. However, trust is of paramount importance in these systems' rise to fame, as was demonstrated in this article. If people do not trust the decisions of the system, why should they use it? Equally important, as the author notes, is the results, as they pertain to how accurate the developers must make the system, before it will be accepted by the population. Of course, AI systems can never be 100% accurate, all the time. So a tradeoff must be reached by the developers. The AI system cannot live for eternity in the researcher's lab, being perfected. But, it can also not be released prematurely, or it will not be trusted, and thus not be accepted. Let us now return to the specifics of the study at hand and consider the methods and predictions.

In the Travel Information Business, it is known that acceptance of the system must be reached rapidly. The more such systems that become installed, the better the information delivered by the system, since it will have more access points from which to query. But, for this increased accuracy to come about, a good number of "guinea pigs" must want the system installed in its relative infancy. This can only happen if they trust the system sufficiently, and do not give up on it as it gains in popularity (and, at the same time, accuracy). Fortunately, the author uses some (now legendary) hypotheses by Muir, which have been supported many times over the years. One such hypothesis is that people generally will trust a system at first, with no rational basis for this trust. However, her next hypothesis is that, not surprisingly, this trust is dynamic, and increases or decreases as they use the system. Finally, she notes that, once trust has been lost, it can be rebuilt, however this process is definitely not trivial. What this means for the development of Travel Information Systems is very clear: the systems have a "grace period" at the start, where people will have trust in the system. But give people too many incorrect decisions, over an extended period of time, and the trust in these systems is essentially gone. With this in mind, Fox had his participants drive in a realistic Highway-Driving simulator, with the goal of avoiding congestion at various forks in the road. One of the two choices would lead to congestion, and the other would lead to clear road. The Traffic Information System would be present, giving advice on which of the two alternatives to take, but would purposefully give incorrect information on some trials. One group would receive information which started off at 40% accuracy, and increased throughout the trials to 100%. Another group would receive 100% accurate information, which declined progressively through the trials, to a final rate of 40%. Unfortunately, the author, as of yet, does not have the results of the study available. The specifics would, no doubt, be intriguing. Which group would have the least trust at the end? Would the initial grace period, for those beginning by getting wrong answers, run out before the accuracy of the

answers increased? Would those who began by receiving highly accurate responses be subject to a primacy effect, effectively nullifying the later misleading suggestions? These answers will have to remain hidden for now. However, since the study is based on such a solid foundation, general principles are not hard to extract. Apart from a “free” introductory period, mistakes made by AI systems correlate with decreased trust, and hence decreased usability.

From a purely goal-oriented viewpoint, then, one key to successful integration of these systems is to make people trust them. The question now, of course, is how to attain this trust. Humans trust them initially, but how can the trust be fostered? Obviously, people may become skeptical (as in the Travel Aide example previously presented) and begin not to believe the system’s conclusions.

Does Human Understanding Help?

For some potential insight, let’s consider a study by Bauhs and Cooke, (1994). These authors wondered whether or not information on a system’s method of development would help users to trust and use the systems more. Let’s informally sample the possibilities. By understanding the techniques that a system uses, perhaps people will be impressed, or overwhelmed, by the intricacies, and be forced to appreciate the power of what he does not understand. However, an alternative may occur. Perhaps people, by learning how the machines work, will begin to feel that there is nothing special about these expert systems, and that they themselves have better reasoning ability than them. This is, of course, the exact opposite of what expert system implementers want to occur. Which of these two possibilities did the authors find? Is leaving the aura of “unknown” or “magic” intact positive or negative? Let us tread carefully through the study, so as not to dismiss the subtleties.

The authors note that there are two reliable measures of trust which have been used in the past. One, already implicated in the Traveller Aide study, is the extent to which the person listens to, or follows the advice of, the expert system. Secondly, the effect of system advice on the person’s self-reported confidence at the task can be assessed. Using these two methods, the author’s tried to gauge the effect of system information on trust of the system. The subjects’ goal was to determine the fastest route between certain locations. Two groups were used—one group given prior system development information, and the other deprived of the information. Furthermore, these groups were subdivided into those who received positive feedback from the system after a trial, and those who received mostly negative feedback.

The process went as follows. The subject told the experimenter what he believed was the fastest route, as well as his confidence in his answer. Then, the subject had to consult the expert system and receive its feedback, also rating his confidence on this advice. Finally, the subject had to make his ultimate decision, and was then given information on how accurate the advice from the expert system had been, regardless of whether or not the subject chose to use it. With the study explained in sufficient detail, let’s survey the important results. The authors found that those in the low reliability condition (I.E. the advice from the expert system was often wrong), began to take the advice of the expert

system less and less, as they progressed through the trials. Based on our analysis up to this point, it can be inferred that they began to trust these systems less. And it is well established what happens in this case (I.E. the users stop using the system). The results from the high reliability condition are even more interesting. Those given system information also showed a pattern of decreased trust throughout the trials, whereas those not given system information did not show this. The authors point out that this could be due to the subjects believing that, based on what they know about the system, they have better reasoning ability, and so should go with their instincts instead of trusting a sub-par system. Finally, looking at the change in confidence of subject's, only those in the System Information group correctly adjusted their confidence level in accordance with the reliability of their system.

Admittedly, the situation looks even more confusing now, after this inconclusive study is considered. Is giving people information about the expert system helpful? In terms of increasing the trust that humans have when dealing with these systems, and hence their tendency to rely on them, the answer is no. But giving system information to humans isn't all bad, as it allowed them to correctly be more or less confident in themselves, depending on system accuracy. It appears, therefore, that continuing to let people think that the system is somehow magic is more effective in making them believe in the system's choices. But is this morally the correct thing to do? Should information be withheld just because it allows expert systems to be increasingly relied upon? What if the system is not entirely accurate, all the time—would it not be more humanistic to inform clients of this, teach them the methods of the system, and allow them to choose whether or not to trust it? This may be better left to the philosophers of our time. For our slightly myopic purpose at hand, system information does not increase the level of trust felt by a human, for an AI system. Because of how necessary trust is, this cannot be seen as a step in the right direction, and other methods will have to be investigated. The AI discipline is relatively quiet on other such avenues, so it is time to carefully branch out to a different area, and apply what can be learned to the current problem.

To recap, the issue at hand is how to increase the level of trust between a human and an AI system, in order to allow the system to reap the benefits of being trusted. Perhaps the problem lies in the type of relationship that researchers are attempting to form between humans and machines? Or maybe, trust is a built-in mechanism of humans, which cannot be extrinsically altered? Are there different types of trust, leading to the possible conclusion that researchers are currently trying to manipulate the wrong type? Solid answers are still lurking in the shadows, however the field of personality psychology does lead us to some clues.

Which form of trust is important?

Specifically, let's concern ourselves with a study of trust done by Couch and Jones, (1997). Trust has been studied very extensively in the literature, and so these authors had the daunting task of trying to find consistencies between somewhat disparate findings. While these specific conclusions are proving elusive, some general statements, fortunately, can be made. The article references a well known scale called the trust

inventory, which breaks trust into three components, which have proven reliable and valid. The first component relates to partner trust, or the extent to which someone trusts their romantic partner. Second is network trust, which measures the amount to which someone trusts their social network of family members and friends. Third is a general type of trust, encompassing a person's general feeling of trust for the rest of humanity. While this may in itself be fascinating, let us keep in mind our purpose, and speculate on why this matters. If trust is, indeed, made up of more than one entity, then which one is important when dealing with AI systems? In other words, which type should the system seek to strengthen? Would partner trust be implicated, and how strongly? Or perhaps, an AI system would be seen as part of a social network, so that network trust would be more pertinent? Are these types of trust even externally malleable? Returning to the present paper the authors indeed found different correlates which predicted the level of various types of trust, in an individual. In general, relational trust was found to be strongly associated with the person's relationship quality. Certainly, then, an improvement in relationship should correspond to an increase in this type of trust. However, although not significantly, global trust seems to be more of a personality trait—a part of the person's makeup. Should global trust be implicated in dealings with AI systems, then there is a problem. Namely, personality traits are, by definition, not trivial to change. A “distrusting” temperament may, then, be enough to deter a person from interactions with AI systems.

Section Summary

At this juncture, let us collect ourselves and paint the complete picture of what is known about trust in the field. As argued, trust is important for AI systems to gain, from the humans operating or relying on them. A lack of trust funnels into a lack of usage, a lack of interest, and overall, a lack of respect. But, as repeatedly encountered, gaining trust is not a simple matter, for several reasons. Firstly, research has not yet found a reliable way of doing this, although some respectable attempts have been made. Secondly, researchers may not yet even know which type, of several, is the mitigating factor in whether people do, or do not, have faith in these systems. There are multiple types of trusts, some more pliable than others. Future research, then, is necessary, to illuminate these issues. It may, and perhaps should, come as a shock that this is the current state of affairs. With all of the innovation shown in part 1 of this paper, it is obvious that AI has grown substantially as an applicable discipline. Yet, at the same time, a central, critical issue has been swept under the carpet, in favor of flashy new methods of integrating AI into society. Trust is a showstopping issue. No matter how far the field progresses, this issue must be tackled sooner or later, or the move toward the usage of AI systems may not be as warmly accepted as researchers may believe.

Is this too bold a statement? Surely, people are more forgiving than this! Let us quiet this criticism before moving on, by briefly sampling writing by Holt (1996). The author declares, as was done earlier in this paper, that quality of a product is paramount in order to gain acceptance and trust. To illustrate, when purchasing baby products, parents check the packaging for signs of credibility (for example, the British Kite sign). They are not willing to let their children interact with products which have not passed strict guidelines,

in order to receive such insignias. Au contraire, the author notes that, in software, no such guarantee is given—people are basically running software at their own risk. Software seems to be received just fine by the general public, right? Although a reasonable claim, consider the fact that the so-called RAID technology has been invented to (as its name implies) provide redundancy for the data. Backup software is used rampantly by corporations (and individuals alike) to safeguard their hard work. Software applications include Crash Recovery, AutoSave, and other security features, to protect the user from their fallibilities. It seems, therefore, that not only do people distrust software, but software distrusts software. But, due to the closeness with which AI systems will interact with us, they must be treated as baby's products, more than they are treated as software. So, they must have some sign—some indication—that they are trustworthy.

This concludes our survey of the trust literature. Let us now concern ourselves with the second critical issue in the bid for AI integration, namely, safety.

Part 3: Safety

From the extensive literature cited so far, it should not be a revelation that safety, security, and related aspects, would be seen as a prerequisite, by the majority of people, before accepting expert systems into their lives. In fact, some of the earlier references will be called upon later on, this time in light of this statement. However, there is something even more basic, more malignant—although less obvious—to consider first. The ultimate problem is, in fact, not that AI systems are safe for humans to interact with, at the current time. To be sure, this should not be minimized, but it can only be achieved through circumventing the larger problem. This problem, in short, is that AI systems may become smarter than humans one day. Once this occurs, if safety is not at the core of how these systems reason, think, and deal with humans, then this is where the real problem could lie. As long as humans are dominant, any AI safety issues can relatively easily be corrected. But when the pendulum of intelligence rotates over to the other side, what are humans to do then? If these systems are not safe at this point, could humans become victims of their own creation?

To see that this is not an overreaction, let us turn to writings by the Singularity Institute for Artificial Intelligence, Inc. (2004). The word Singularity, in the organization's title, refers to the creation of technology which has greater intelligence than humans. They note several branches of computer science which are interested in achieving this long-term goal, and, of course, one such discipline is Artificial Intelligence. Over the centuries, humans have become rather accustomed to being the most advanced species on the planet, when it comes to cognitive ability. In fact, many people may find it laughable that any system made by humans could ever become more intelligent than their creator. Just like a scholar student can eventually become more adept than his teachers, so too, it will be argued, can a technological system become more intelligent than humans.

According to this organization, all that is necessary for this to happen is for systems to have the ability to enhance their own intelligence. A miniature version of this can be seen

in the ability of programs to modify their own source code. But once they have the ability to enhance intelligence, why is it certain that they will outrun humans in the race to greater and greater skill? The answer to this remark is also trivial. It lies in the fact that human evolution is agonizingly slow, whereas technological speed doubles every year or two. Coupled with this is the fact that transmission speeds through a human neuron are millions of times slower than transmission through computer circuits. Not only are they going to be more intelligent than us, but they'll make their decisions faster!

With this said, let us go back and quantify our bold statements on the necessity for systems to be reliable and safe for humans to use. The Singularity Institute makes it clear that, at the field's current state, AI can not really harm, or help, humanity. This seems like an unsubstantiated claim based on what has been discussed so far, but let us leave them off the hook for the most part, and persevere with our argument. Assuming that AI cannot hurt people at its current state, why is safety so critical? Surely, obtaining trust—our last topic of discussion—is more immediately necessary? We now know that this thinking is correct. Leaving safety concerns on the backburner, for too long, may cause a too-little-too-late effect. If the systems become too advanced, prior to safety concerns being engrained in their chips, safety may never become part of their lifestyle. Relying on cliché, “you can't teach an old dog new tricks”. So, safety concerns are of utmost importance now, today, and not in years or decades from now. Machines can't hurt us now, and implementers should be using this time to ensure that they do not hurt us in the future.

Force: The Next Frontier in Feedback Mechanisms

Before resampling the data in section 1 as promised, let's look at one new area where safety is involved. Specifically, the work to be analysed is from Iwata (1990). This author was studying the field known as Artificial Reality, which concerns itself with appealing to various human senses in order to make a virtual space appear as real as possible. As noted, this is useful for the remote manipulation of objects. Certain senses have, throughout the short life of the field, been given priority to others, and have hence been studied more. For example, consider the senses of sight and hearing. Computer video cards alone have as much processing power as a CPU these days! And audio is not far behind—with technologies like surround sound, audio has undertaken a new level of realism. But, what happened, the authors ask, to tactile feedback? This has evidently collected some dust over the years, but is at least sampled in this paper. It is important, if for no other reason than to narrow the gap even further between being virtual, and being real. The solution put forth, presently, is to rely on force-feedback for tactile responses. Forces, specifically in this research, are applied to the operator's palm and hand. The rest of the article is somewhat technical, but the main observation, and topic of discussion, has already been extracted. Namely, systems will soon be interacting directly with people, not through safe mediums such as graphics and sound. These systems will actually be coming into contact with humans and, while this small force feedback situation is safe enough, it does pave the way to greater potential. At the extreme is when so-called service robots are interacting with people; for example, as assistants to elderly or otherwise disabled individuals. Surely, the force they produce must be gauged

correctly, and monitored according to who the robot is working with, otherwise safety issues begin to abound.

Safety in the Hospital

At last, let us return, with brevity this time, to our initial analysis of the motivational aspects of AI, but now with a weary eye focused on safety. There is little choice of where to begin—the work on robot-assisted surgery is certainly the most salient topic (Theodossy & Bamber, 2003). There is a further roadblock, not previously addressed, that we can consider now. Succinctly, these devices are meeting resistance by such organizations as America's Food and Drug Administration (FDA) because they are a potential risk factor. Luckily, the article also provides some mechanisms whereby safety can be increased, and it is these which we will concern ourselves. Firstly, these robots can be restricted, by the types of movements they are allowed to perform, and the range in which they are allowed to move. In other words, precautions can be instated to minimize the chances of the robot causing any damage. Perhaps this goes against standard practice in computer science—namely, the fact that, often, general routines are better than specific ones. But, robot specificity may be different. If the robot doesn't have the ability to do anything unsafe, then all of its actions must be acceptable. Surely this is a step in the right direction. Another technique mentioned is to make these robots passive, versus active—a distinction raised in section 1 of the paper. Although, it was argued, active robots do have their purpose, and may be more powerful than passive ones, power should, perhaps, be sacrificed for control. All-powerful robots would be nice, but if this ends up being a safety hazard, then it is definitely not desirable. This last point ties in nicely with the aforementioned concept of Singularity. Perhaps the key to successful utilization of AI systems is not to let them “run away with their intelligence”. Instead, the idea may be to limit what these robots can do, keeping humans in ultimate charge at all times. This is, evidently, a delicate balance. Too much power, as we've seen, leads to potentially unsafe machines. Too little power and control, and legions of untapped potential may be pounded into submission and represent little more than the personal computers of today.

Museum Madness

Perhaps at this point, the reader envisions a simple solution to the previous conundrum. After all, doesn't the problem lie in the criticality of the procedure that the robot is trying to perform? Why not simply keep robots out of the surgery room, and thereby eliminate the problem altogether? This, perhaps, sounds convincing and lucid, however it is crucially flawed. It will be argued that the surgery, per se, is only one representation of a deeper, underlying problem. It will be argued that, any time there is the possibility of robot-human interaction (even when it does not come under the knife), there is cause for concern. To see this, let us take the example of the tele-presence museum study (Agah, Tanie, Ohkawa, & Iwamoto, 1997). Imagine that this robot was wandering around a full-sized museum, heavily populated with humans. Even with this simple scenario, safety concerns must be appreciated. As the authors note, the robot must be designed so as not to harm anyone in the museum, or damage any object in the museum. This is complicated by the fact that this robot uses human intentions to guide its actions. This could

potentially cause a conflict between what the robot is asked to do, and the safety mechanisms that the robot has built-in. The extreme of this situation, of course, occurs if a nefarious user is allowed to interact with the robot, but, in daily operation, human-guided accidents, or mistakes, can surely occur.

This raises yet another paradox for the AI system to deal with. While in a conflict, which route—machine logic, or human intention—should be followed? Should humans be disobeyed if their commands are not in agreement with the AI program of operation? If this is so, perhaps this is just an attractive façade covering the Singularity problem previously discussed? Should they then allow human intervention to override their internal logic? It will be left for researchers to determine its solution. Considering the study one last time, there is one safety aspect which has remained hidden throughout our discussion so far: the safety of the robot itself. These devices will certainly cost millions to make (at their inception, anyway). Furthermore, if the robot ends up in an unsafe environment, it may have a ripple effect and cause other parts of the environment to follow suit. Consider the museum robot being subjected to water damage, or electric shock. Monetary considerations aside, this is also detrimental to any people in the area of the robot at this time, or the museum structure itself. Therefore, it is obvious that robot safety acts as a feedback loop, affecting human safety if it is compromised. The safety of the machine itself may ultimately be as important as the safety of those who surround it.

Section Summary

It is now obvious that, because of the future roles of AI in our society, safety cannot be overlooked. The situation looks bleak, though, since there appear to be so many things wrong with the current state of AI in terms of safety and (as previously argued) trust. Isn't there any good news that can be taken from this, before we tackle our final topic? Can't AI systems prove at least partly useful to use in these two areas?

Interlude: Expert Systems Assessing Trust

To leave on a positive note, consider work by Chadwick and Basden, (2001). This paper deals with the concept of the Public Key Infrastructure, which will be described briefly. When an electronic message is received, one cannot be certain that it originated from the indicated sender. To allow this match to be made, the idea of public keys was introduced. These keys are assigned to individual people, so that, when a message is received encrypted with this key, the recipient can really know who it was sent from. A Certification Authority (CA) is the entity which provides the binding between individual and key. Of course, this is way too optimistic to work in practice, and not even keys are totally secure. People should only trust these keys to an extent, and not believe as Gosel truth, that a message is genuine. Could expert systems help in evaluating the amount of trust a key deserves? It looks as though this is the case! These authors developed an expert system which worked by asking the recipient several questions about the CA, and returning a number from 0 to 1, assessing the relative amount of trust that could be placed in it. The authors admit that not all factors could possibly be considered, and so the return value from the system cannot be as precise as would be desired. But, at the very least, this study shows it is possible for AI systems to assess trust (a subjective, humanistic quality)

with objective data. This is certainly a bit of good news. Perhaps, if AI systems could assess what they are doing, as a function of recipient trust, they could calibrate their actions in accordance with what humans are receptive to? The idea is in its infancy at this point, but it does reveal a potential avenue of future research. With this, let us move on to our last issue of discussion.

Part 4: AI as Teacher

This final point of interest is a culmination of much of what has been discussed thus far. The implications of being taught by an expert system are important, since they directly affect how children of the future will learn. Are AI teachers more effective, more credible, or just somehow “better” than the teachers of today? Even if they are, some merit must be attributed to having a live person teaching society’s children. But can’t this “liveness” be complemented or enhanced by accessory AI systems? These types of circular arguments, as we’ve come across before, are not easy to solve. But there are certain things, all of which have been previously discussed, that AI teachers must do, to have any chance of survival in our society. Of particular interest is their effectiveness in presenting and testing material. In fact, due to potential negative pressure on this new paradigm, they may require effective teaching abilities that far outdo what is common practice today. Second, the trust factor raises its eyebrow at us, once again. Students, parents, and those with the power to implement such systems, must fully trust the information and the methods of the system. There is no point doling out AI teachers with software bugs, or occasional malfunctions. Education is basically sacred to our society, and is something not to be sacrificed in our youth. The argument, then, is that these systems have to be extremely trustworthy before they will even be given serious attention. The issues discussed presently are critically important, however do not appear to be at the forefront of today’s research. This is perhaps due to the relative novelty of the field, and so we will follow suit and discuss some paradigms which have proven successful, at least at this early stage.

Foreign Language Learning

Let’s begin with some fascinating work by Micarelli and Boylan, (1997), which focused on the learning of a foreign language. This is evidently important in the home country of these authors (Italy), but equally important in bilingual countries like Canada and the US. Effective techniques would surely be embraced by many, and it is to one specific technique outlined in this paper that we now turn. The technique is known as Conversational Rebuilding (CR), and differs at a very low level from conventional techniques. The authors note that standard practice is for a foreign language instructor to ask for individual utterances, and to label them as completely correct, or totally wrong. In other words, there is no middle ground—if the word is not spoken as a fluent speaker would have said it, it is not deemed acceptable. Perhaps this is not the most appropriate way to go? CR, by contrast, allows “degrees of correctness” when learning the new language. It is not practical—some may argue, not possible—to get a foreign tongue correct the first time. However, something that vaguely sounds like the foreign language is better than something which doesn’t resemble it at all. The ultimate goal is

understandability—if the listener can understand what the learner is getting across, it is a success.

In CR, one common technique, then, is for the teacher to begin with utterances that the learner is familiar with, and gradually transform them into proper utterances. The learner is never flat-out wrong, and this is used to motivate the learner to get closer and closer to actual speaking of the new language. The authors label this transition period as interlingual, or, literally, between languages. Now that the framework has been laid out, how can an AI system help? The present authors have designed a system which facilitates the learning of the foreign language using this technique, because, they argue, it is readily applicable to the sorts of things computers are good at. Namely, their system allows the learner to begin with their version of the new language as a starting point. Then, as long as it can find some semantic basis—some similarity—to the new language, the process of CR can begin. The authors argue that this is reason for optimism, because it shows that a novel technique can successfully be implemented on a machine. The crux of the work that must be done by the AI system appears to be an application of pattern matching. Using this technique, this is effectively what teachers do—they pattern match the learner's utterance against the words which most resemble it, and take a guess as to what the intended meaning was. A detailed example of this is given. A teacher may show a film to her students, and stop it at the point in which a character is about to say something. She'll then ask the students to determine what will be said next. The students respond, and the teacher, focusing on one inconsistency after another, eventually guides the students to a sentence which conforms (exactly, or just related by meaning) with the following utterance in the film. The authors point to the teacher's thought process, of which mistake to correct at which point, as the most difficult to emulate in the AI system. One other shortcut, if we dare call it that, is that the AI system accepts only one answer as correct, and not other answers which, although different, would still be acceptable in that situation. While seemingly restrictive, the authors give reasons why this technique may be better than a variable-answer technique. For example, they state that the variable technique could disorient students, since there is no single correct answer. Now, moving to the results, how do students feel about the system? The conclusion drawn in the paper is that, while learners may not like using the system more than learning from a live teacher using CR, they would prefer to use the system over a teacher using different (fixed-point) methods. This also speaks to the possibility of incorporating such systems into the learning process as only an accessory, not a replacement, to the teacher. As the authors point out, there are times when their system has no idea of how to respond to an utterance which is very wrong. Perhaps in these situations especially, a human teacher would be more effective, and respond more appropriately.

The results of the present study are very important, and should be understood fully. The claim is that AI systems are not ideally intended to do any replacing or removal of today's teachers. This, as just realized, would not be beneficial to students. Instead, consider how teachers use overhead slides, laptop computers, and other technological aides, to bring their lectures to life. To a degree, it appears that AI systems may fall more into this category of tools. It must be emphasized that this is not a bad thing. The power of AI should be harnessed to do what it can perform most effectively, and be most helpful

to those using it. Hopefully, one day, AI systems will work alongside teachers to bring a new, interactive dimension to academic life.

From Pascal to C

While the previous example proved highly illustrative, it is not enough to subsume the venerable field of AI-instructibility. To continue our look, then, let us turn to work by Fix and Wiedenbeck, (1996), and see how AI systems can help computer scientists, no less. The motivation for the work was found in a common misconception dealing with programming languages. The argument is that, once someone learns one programming language, others will be easily understood since, after all, they only differ in syntax. This claim is refuted, and several examples are given as to why. While it may be true that languages are syntactically different, they are also conceptually dissimilar. Therefore, a programmer could easily convert from, say, C to Java, with little difficulty. However, the new program would not take advantage of features that are only found in Java, and not C, because the programmer is not familiar with Java. This leads to inefficiency, lack of understandability, and overall, time wasted, as the program does little more if it cannot take advantage of the purpose of the conversion!

However, in our fast-paced society, programmers may not have time to learn a new language from the ground up, as was done for their introductory language. If only there was a way for programmers to attempt the new language, while at the same time, being instructed and prodded in the right direction, to take advantage of the new methodologies and tools available. This should sound eerily similar to the arguments presented above dealing with foreign language learning. Therefore, it is no surprise that an AI system for helping the programmer transition between languages, could be devised. After all, it is once again a problem of pattern matching, which, it is known, is prime territory for an AI instructor to jump in.

These are the take-home details of the system developed by Fix and Wiedenbeck, but for completeness and interest's sake, we persevere into some specifics. The language in question is known as ADA and, for reasons unimportant to us, is conceptually different than traditional languages like C and Pascal. Therefore, it may require that a student's "plan of attack" for a problem change. However, there are several strategies that work in all programming languages, including ADA. So, the main goal of the AI system is to help the programmer understand when his plan must change, and when it may remain intact. Note how this is completely different from training a novice programmer, who effectively has no previous plans to rely on, and so cannot make the mistake of using them. Paradoxically, it looks as though previous experience can have a small negative effect, which, by developing this AI system, is hoped to be eradicated.

The operation of the system is, again, similar to the foreign language example. The system presents the learner with a function which it would like written in Ada. The learner begins with a template, and replaces its lines with Ada code. While doing so, he will undoubtedly make mistakes, and this is where the AI system becomes helpful. If the system notices a mistake that, say, a Pascal programmer would make, it informs the user of this, and inserts the missing code so that the user will know for next time. Similarities

continue to abound, as this is effectively a reincarnation of Communication Rebuilding (or, more appropriately, Programming-Language Rebuilding). The goal is to take the programmer, one step at a time, from a Pascal or C way of thinking, to an Ada way of thinking. However, this article does make clear some points which have not been duplicated in the previous example. Specifically, the idea of patience in this example is key. The transition between programming languages could very well be a long, arduous, and awfully boring experience. After all, the experienced programmer may not want to “waste time” learning something he was taught years ago, only in a different context. With this in mind, the teacher of the system must be patient and allow the process to proceed at the rate which is most comfortable for the learner. Arguably, there is no better match for these requirements than a computer. While human teachers get tired, run out of time, or even become discouraged, the AI system feels none of this.

Let us finish with some memorable examples from Woolf and Hall, (1995). The authors describe a system called the Cardiac Tutor, for teaching medical students about cardiac resuscitation. The information that the system uses was obtained from known procedures for working with the heart and, not surprisingly, is implemented as a knowledge base. Students who use the system are shown the display of a medical room and have to save the patient inside. The system provides all necessary information about the patient, as well as clues to assist the student in his job. The remarkable fact is that two classes of students have said that this method of learning was just as effective as working one-on-one with an instructor. Also, it is stated that the use of AI systems based on knowledge bases allows students to master material in one-third the time it would take using conventional techniques. What is the reason for this? The authors speculate that AI systems are better at fine-tuning themselves for the individual student they are dealing with. By zeroing in on the specific difficulties, they can be realized and overcome more quickly. However, knowledge-base systems do have their difficulties, including finding suitable representations for knowledge, and assessing the errors that students are making.

Section Summary

Overall, what can be learned from the two previous powerful examples? Firstly, there is evidence showing that students do enjoy—even condone—AI system’s interjecting into their educational careers. Second, it is obvious that, at least in some specific cases, AI systems can prove effective teachers. More research is necessary to determine how valid the claims in these two examples are, however the big picture is clear. AI teachers will, some day, be a definite asset to students and others wishing to become educated. It opens up the possibility of true learning from home, student-guided lessons, and teaching ability which doesn’t lag far behind real teachers. There are issues to solve first, as we’ve seen, but which disciplines do not have their caveats? The waters ahead do have some rapids, but it is hoped that the reader can envision the future—a future consisting of robots in the classroom. Keep in mind, as was said at this sections beginning, that the goal of these systems is not to replace real teachers. In some ways they compliment a teacher’s ability, and in other ways, they can teach a student on their own, and be very effective at doing it.

Concluding Remarks

Let us finish our tour of AI integration into human society by first summarizing our progression through the field, and then speaking in general of the main principles extracted.

We began by abashedly stating that AI, now and in the future, will be of great importance and usefulness to the general public. Many examples were presented to back up this claim. For example, it was shown that AI systems are useful as walking aides, as entertainers, as government assistants, at communicating, at responding to individual differences, and so on. Therefore, with all of these possibilities, it is not far-fetched to agree with the plausibility of using AI systems to benefit many people. Then, however, we moved into some issues which clouded our optimism for the time being. AI systems, it was argued, must obtain the trust of those using them, before they can be successful. It was shown that this is a nontrivial task, and work from various areas of psychology was sampled to try to find some solutions. As if this wasn't enough, the necessity for systems to be safe and secure was also exposed. Being safe is also no walk in the park, as the memorable example of the surgical procedures would testify. Then, to complete our survey, the potential of AI systems acting as teachers was discussed. It was found that the most appropriate way for these systems to help those in schools, was to act as an aide for current teachers to use, although sometimes, these systems appear to have the ability to teach on their own. So, there are all of these wonderful possibilities for AI systems, but several hurdles in the way.

Throughout the paper, a conscious effort was made to give humble suggestions on how certain problems which arose could be solved. For example, by making robots passive, and not active, several caveats dealing with hospital care could be reduced. By improving the communication links between human and robot, AI systems could better understand what the human is requesting. Finally, by understanding which type of trust is implicated in human-machine interactions, researchers could work on methods for having the level of this trust improved. All of this was determined by simply perusing research that has already been done, in various fields of computer science (robotics, security) and psychology (evolutionary, personality). Granted, there are implementation details that are much more difficult to solve than these. But the point is that it appears that researchers are well on their way to achieve the goal of coexistence of humans and machines in the same living space.

The task may appear daunting now, because the apex of this coexistence will not occur in the immediately impending future. However, consider the relative ease with which personal computers have invaded our households. Within a span of 20 years, hundreds of millions went from envisioning a computer as being hundreds of pounds heavy and living in a freezer, to a device which many could no longer live without. It will be argued that the fate of human-AI interaction will occur along the same lines. The future is definitely bright. The current problems will be solved sooner or later, and probably give insight to researchers in the meantime. It is no joke to begin the introduction of AI systems into our houses, and much time should be spent on their design. The current issues will help in

slowing down the process, giving everyone a chance to think about the ramifications. It is fascinating, given the knowledge obtained in this paper, to consider how much of what the reader has just read will be obsolete in a short time. We are in a technological explosion, with thousands of minds working towards common goals. One ultimate goal of much AI research is to have safe, reliable, trustworthy, and helpful relationships occurring among humans and machines. The scientific community will no doubt devour these challenges, and it is only time until the motivations of integration, discussed in section 1 of the paper, come to fruition.

References

- Agah, A., Tanie, K., Ohkawa, K., & Iwamoto, K. (1997). Tele-Museum: Multimedia Interface and Control for Exploring a Remote Museum from Home. *Proceedings of 6th IEEE International Workshop on Robot and Human Communication*, 448-453.
- Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the “tragedy of the commons”. *Evolution and Human Behavior*, 25(4), 209-220.
- Bauhs, J., & Cooke, N. (1994). Is knowing more really better? *Conference on Human Factors in Computing Systems*, 99-100.
- Chadwick, D., & Basden, A. (2001). Evaluating Trust in a Public Key Certification Authority. *Computers and Security*, 20(7), 592-611.
- Couch, L., & Jones, W. (1997). Measuring levels of trust. *Journal of Research in Personality*, 31, 319-336.
- Dautenhahn, K. (1995). Getting to know each other - artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16, 333-356.
- Fix, V., & Wiedenbeck, S. (1996). *Computers and Education*, 27(2), 71-83.
- Fox, J. (1996). The Effects of Information Accuracy on User Trust and Compliance. CHI Conference Companion 1996, 35-36.
- Holt, J.D. (1996). Practical Issues in the Design of AI Software. *Engineering Applications of Artificial Intelligence*, 9(4), 1996.
- Iwano, Y., Sukegawa, H., Kasahara, Y., & Shirai, K. (1995). Extraction of speaker's feeling using facial image and speech. *Proceedings of the Fourth IEEE International Workshop on Robot and Human Communication*, 101-106.
- Iwata, H. (1990). Artificial reality with force-feedback: development of desktop virtual space with compact master manipulator. *Computer Graphics*, 24, 165-170.
- Kawamura, S., Ida, M., Wada, T., & Wu, J.L. (1995). Development of a virtual sports machine using a wire drive system - a trial of virtual tennis. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 111-116.
- Kobayashi, H., Une., T., Takahashi, S., Mikuriya, K., Ohya, A., & Yanagawa, M. (1994). An autonomous eye robot for tele-watching. *Proceedings of 3rd IEEE Workshop on Robot and Human Communication*, 323-326.
- Komura, T., Kuroda, A., & Shinagawa, Y. (2002). NiceMeetVR: facing professional baseball pitchers in the virtual batting cage. *Proceedings of the 2002 ACM symposium on Applied computing*, 2002.

Langer, E., & Rodin, J. (1976). The effects of enhanced personal responsibility for the aged: A field experiment in an Institutional Setting. *Journal of Personality and Social Psychology*, 34, 191-198.

Micarelli, A., & Boylan, P. (1997). Conversation rebuilding: from the foreign language classroom to implementation in an intelligent tutoring system. *Computers and Education*, 29(4), 163-180.

Mori, H., Kotani, S., & Kiyohiro, N. (1994). Human interface of a robotic travel aid. *Proceedings of the 3rd IEEE International Workshop on Robot and Human Communication*, 90-94.

Oskamp, A., & Lauritsen, M. (2002). AI in Law Practice? So far, not much. *Artificial Intelligence and Law*, 10(4), 227-236.

Singularity Institute for Artificial Intelligence. (2004). Retrieved February 14, 2004, from <http://www.singinst.org>.

Theodossy, T., & Bamber, A. (2003). Model Surgery With a Passive Robot Arm for Orthognathic Surgery Planning. *International Journal of Oral and Maxillofacial Surgery*, 61(11), 1310-1317.

Woolf, B., & Hall, W. (1995). Multimedia Pedagogues. *Computer*, 28(5), 74-80.

Zhou, Z. (2003). Three perspectives of data mining. *Artificial Intelligence*, 143(1), 139-146.