Teaching Artificial Intelligence in Introductory Cognitive Science Courses

Sara Owsley Sood
Pomona College CS
185 East Sixth Street
Claremont, CA 91711
sara@cs.pomona.edu

Abstract

At the intersection of Artificial Intelligence and Cognitive Science is the question "What is intelligence?" The answer to this question has great impact on the philosophical questions in AI: "Can machines think?" and "Can machines feel?" In this article, I outline how to approach these questions in teaching. I will provide a survey of topics and resources intended for use in teaching artificial intelligence in the confines of an introductory cognitive science course. In the first part of the article, I will address the philosophical roots of Artificial Intelligence, surrounding the question "Can machines think?" In the second part, I'll discuss the question "Can machines feel?" by introducing the notion of emotional intelligence and describing recent systems that embody this idea.

The Philosophical Roots of Artificial Intelligence

A standard definition of Artificial Intelligence is building machines that think or act humanly or rationally (Russell, 1995). This definition is necessarily broad, accounting for the different schools of thought that surround the term *intelligence*. At the base of the field of Artificial Intelligence lie the questions: "What is intelligence?" and "Can machines think?" Much debate and discussion in the history of the field has focused on these questions, including seminal works by Alan Turing and John Searle. In the section that follows I'll provide resources, topics, assignments and discussion questions intended to guide students through the historical foundations of the field of Artificial Intelligence.

In leading a class through an exploration of these topics, certain readings are critical:

1) Alan Turing. "Computing Machinery and Intelligence," Mind, New Series, Vol. 59, No. 236 (Oct., 1950), pp. 433–460.
2) John Searle. "Minds, Brains, and Programs," Behavioral and Brain Sciences, Vol. 3, (1980), 417–424.
3) Robert French, "The Chinese Room: Just Say 'No!'," Proceedings of the 22nd Annual Cognitive Science Society Conference (2000), pp. 657–662.

Alan Turing

The first major work that should be examined is Turing's 1950 paper titled "Computing Machinery and Intelligence." This paper did more than join the discussion of what it means to be intelligent, what it means for a machine to be intelligent; it laid the groundwork for what would be the goals for the field of Artificial Intelligence for decades to come.

Turing's paper restated the question "Can machines think?" by describing a behavioral test for intelligence, the "Turing Test." The Turing Test, modeled after a party game called the imitation game, was intended to test whether or not a machine was *intelligent* by testing if it could simulate intelligent human behavior. The test is set up such that the system is in one room, a human in another, and a human interrogator in the third. Through passing questions to and receiving answers from each room, the interrogator's goal is to determine which room contains the machine. The goal of the machine, and the test of the machine's intelligence, is to trick the interrogator into thinking that it is the human.

In addition to creating the "Turing Test" and addressing objections to such a test, Turing also made predictions for the future. He was optimistic, predicting that by the year 2000, systems could be created

to pass the Turing Test in a 5-minute conversation with a human, 30% of the time. This prediction triggered the decade of optimism that filled the field of artificial intelligence from 1950 to 1960. Turing also predicted that the phrase "thinking machine" would not be a contradiction and that "machine learning" would be an important area of study in the future. The latter is widely known as the most astute of Turing's predictions.

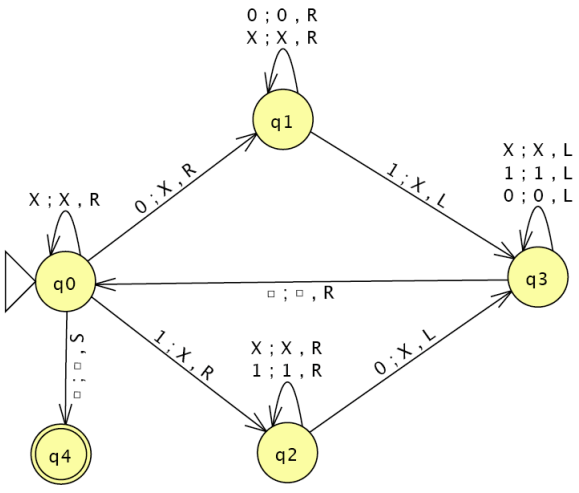
To understand the current state of AI research, it is important that students understand that no machine has passed the Turing Test, as of 2008. The Loebner Prize includes two separate annual prizes: one to the machine that acts most humanlike, and another to the machine that passes the Turing Test (an award that has never been given). Jabberwacky is a chatbot that won the "humanlike" Loebner prize from 2003 to 2006; it is available online and is a nice classroom demonstration for students (Jabberwacky, 2008). Other systems including ELIZA (Weizenbaum, 1966; ELIZA, 2008) are interesting comparisons to Jabberwacky. While ELIZA identifies patterns in the input from the user and matches them to rules in templates provided by a programmer, Jabberwacky learns from talking to users - expanding knowledge on its own. Both systems have online versions that students can interact with (ELIZA, 2008; Jabberwacky, 2008).

An interesting exercise is to have students write out patterns of inputs and outputs to simulate how they would build two chatbots – one that is domain specific (that could talk about movies for example) and one that is general purpose. After completing this task, I typically have the students reflect on the challenges of this task. Having students create rules for both a domain specific chatbot and a general chatbot enables them to understand that limiting the chatbot to a finite space of topics makes the task much more feasible.

An important objection arises when discussing the chatbots described above. Many students feel as though they could be convinced that Jabberwacky was a human. However, there is a difference between talking to Jabberwacky - becoming pleasantly surprised in something intelligent that it says, versus passing the Turing Test. The Turing Test requires that an interrogator actively try to seek out which participant is the machine. This distinction makes the test much more challenging.

The attempts to pass the Turing Test have been unsuccessful for a variety of reasons. French outlined some common failures (French, 2000); he examined the aspects of the human experience that could not be finitely represented. As an example of such a human experience, he explains that machines could never be equipped to answer all possible questions like: "Does holding a gulp of Coca-Cola in your mouth feel more like having pins-and-needles in your foot or having cold water poured on your head?"

Turing's Test was intended to answer the question "Can machines think?" The test itself gave an explanation of what it means to think. However, another important question is "What is a machine?" In his paper, Turing describes a machine as a digital computer – again restating the question "Can machines think?" as "Can a digital computer pass the Turing Test?" Turing's past work actually established an even more precise definition of a machine, a "Turing Machine." The Turing Machine, created in 1937, was intended as a simple theoretical model of a machine that could handle any computation that is possible on digital computers. Teaching Turing Machines in lecture is necessary, but in order for students to truly understand them, I think they must build them. For that reason, using a tool such as JFLAP (JFLAP, 2008) to allow students to easily build, run, and test Turing Machines is quite helpful. Below is an example Turing Machine created in JFLAP. This Turing Machine accepts all strings or words that contain an equal number of 0's and 1's.



John Searle

The next important work to consider is John Searle's 1980 paper titled "Minds, Brains, and Programs" in which he describes a thought experiment called the Chinese Room. In this room, a person (who does not speak Chinese) sits with a stack of papers containing mappings and rules. A person outside the room, a native Chinese speaker, passes a note through a hole in the wall. The person in the room takes this note and follows the instructions in his stack of papers, returning a new note through the hole in the wall. From the outside of the room, it appears that the person inside the room speaks/understands Chinese. However, it turns out that he is just following the rules in his stacks of papers, which provide an intelligent response to every possible input. Searle argues that this person does not understand Chinese and that the entire system (the person, stacks or paper, etc) does not understand Chinese. Since this room is designed to model a Turing Machine, Searle claims that a machine can never truly understand or think. The paper also includes many replies to these claims, as well as Searle's response to each reply.

In describing his Chinese Room, Searle also points out two distinct schools of thought surrounding this topic. The Strong AI side argues that it will be possible to build machines that think while the Weak AI side believes that machines are powerful in many problem-solving tasks, but we'll never be able to create one that handles all cognitive tasks. An interesting exercise is for students is to explore some famous AI researchers (perhaps from http://www.aaai.org/AITopics/) and hypothesize, based on their work, which researchers stand on either side of the debate.

An important follow-up to the Searle reading is French's paper from 2000 titled "The Chinese Room: Just Say "No!". In this paper, French argues against the Chinese Room, not for any of the reasons addressed in the "replies", but instead because he says it would never be possible for a non-native Chinese speaker to perform this task. He argues that the mappings/rules that must be provided are not finite in number. For example, the Chinese speaker could ask about a visceral reaction to a made up word - a question that a native-Chinese speaker could handle. It is not possible to provide all such question/answer mappings, and for that reason, French argues that the Chinese Room should not be considered.

Guiding students through the three readings listed above, while also giving them experience building Turing Machines, examining systems that have attempted to pass the Turing Test, and making distinctions between various schools of thought are exercises that are intended to create a greater understanding of the philosophical roots of artificial intelligence, in particular, addressing the question "Can machines think?"

Emotional Intelligence

After decades of work towards creating *artificial intelligence*, some researchers are now attempting to create machines that are *emotionally intelligent*, finally addressing the question "Can machines feel?" The push towards *emotionally intelligent* machines is an exciting addition to this field, in the direction of machines that truly "act humanly." Many researchers in the fields of Artificial Intelligence and Human Computer Interaction have projected that the machines or intelligent agents of the future must connect on an emotional level with their users (Norman, 2004). This is based on the notion that an intelligent and successful human is not only strong in mathematical, verbal and logical reasoning, but is able to connect with other people. Much recent work in this area has focused on empowering agents with the ability to both detect emotion via verbal, non-verbal, and textual cues, and also express emotion through speech and gesture. In the sections that follow, I will present evidence of this movement towards systems with emotional intelligence while also showing why and how this topic can be introduced in introductory cognitive sciences courses.

The concept of *Emotional Intelligence* became prominent in the late 1980's; however, Thorndike discussed a similar concept called *social intelligence* much earlier, in 1920 (Thorndike, 1920). While one's *social intelligence* is typically defined by their "ability to understand and manage other people, and to engage in adaptive social interactions" (Kihlstrom, 2000); *emotional intelligence* deals specifically with one's ability to perceive, understand, manage, and express emotion within oneself and in dealing with others (Salovey, 1990). Salovey and Mayer define five domains critical to *emotional*

intelligence: knowing one's emotions, managing emotions, motivating oneself, recognizing emotions in others, and handling relationships. A common measure of Emotional Intelligence is EQ (emotional intelligence quotient), as gauged by a myriad of widely published EQ tests.

In the late 1990s, many AI and HCI researchers began to take the notion of emotion and emotional intelligence quite seriously. The Affective Computing Lab within the MIT Media Lab was founded by Rosalind Picard following the publishing of her 1997 book titled "Affective Computing," in which she laid out the framework for building machines with *emotional intelligence* (Picard, 1997). Picard, along with many other researchers in this space, has built machines that can both detect, handle, understand and express emotions.

Before discussing the theories and applications of machines that are emotionally intelligent, it is important to first understand that emotion is an important aspect of intelligence. The evidence that necessitates this work comes from a few different fields. I'll provide references to some of this evidence prior to moving into the work that has been completed in building machines that are emotionally intelligent.

> Is "emotional intelligence" a contradiction in terms? One tradition in Western thought has viewed emotions as disorganized interruptions of mental activity, so potentially disruptive that they must be controlled.
>
> Salovey and Mayer, 1990

As Salovey and Mayer express in the quote above, the common notion is that emotion is a hindrance to intelligent thought. Much work in the field of affective neuroscience has provided empirical evidence that this is not the case, and indeed the opposite is true. Affective neuroscience is the study of the processing of emotions within the human brain. Researchers in this field have shown that emotion plays a crucial role in problem solving and other cognitive tasks within the brain (Damasio, 1994).

In addition to the evidence that arises from affective neuroscience, an even larger body of evidence comes from psychology, arguing that emotional intelligence is critical to a person's success in many aspects of life (Gardner, 1993; Goleman, 1997). For example, through behavioral studies, Bower has shown that mood has a strong influence on memory and social interactions (Bower, 1981).

Models of Emotion

Prior to understanding efforts in this space, one must have an understanding of the various models of emotion that are incorporated into systems. The choice of model is completely dependent on the task at hand; namely what dimensions of emotion can be gleaned from the available input signal, what model lends itself best to internal reasoning within a system, and what type of emotional expression the system aims to accomplish.

The simplest model is one of *valence* (positive or negative) and *intensity*, where sentiment is represented as a single score between -1 and +1, where -1 denotes the most intense negativity and +1 corresponds to the most intense positive score. A slightly more complex model adds the dimension of dominance (a scale from submissive to dominant). In this model, the intensity dimension is called "arousal" (a scale from calm to excited). This more complex model is commonly known as the VAD model, which stands for valence, arousal, and dominance (or PAD where valence is replaced by the synonym "pleasure") (Bradley, 1999; Mehrabian, 1996). This model is commonly used in measuring emotional reactions in humans as these dimensions lend themselves well to this task.

A more commonly known model is Ekman's six emotions model – happiness, sadness, anger, fear, surprise and disgust (Ekman 2003). This six dimensional model is intended to characterize emotional facial expressions and is typically used in systems that intend to express emotion in interaction with users. A mapping between the VAD and Ekman models exists in order to facilitate building systems that both detect and express emotion. For example, a low valence, high arousal, and low dominance VAD score maps to fear in the Ekman model, whereas low valence, high arousal and high dominance maps to anger (Liu, 2003).

Machines With Emotional Intelligence

With an established need for work in this area, I can now present the ways in which researchers are making efforts towards building emotionally intelligent machines. There are various models/definitions of emotional intelligence, but they all boil down to the ability to connect with others by detecting, expressing, managing and understanding the emotions of oneself and others. Efforts in building machines that are emotionally intelligent center around a few key efforts: empowering the machine to detect emotion, enabling the machine to express emotion, and finally, embodying the machine in a virtual or physical way. Projects that incorporate all of these aspects also require the additional ability to handle and maintain an emotional interaction with a user, a large added complexity. In the sections that follow, I will provide examples of systems that approach these tasks – examples that are intended to expose students to work in the space of emotional intelligence.

Detecting Emotion

Work in the space of automated approaches to detecting emotion has focused on many different inputs including verbal cues, non-verbal cues including gestures and facial expressions, bodily signals such as skin conductivity as well as textual information. The end goal in building systems that are able to detect an emotional response from a user, is to handle/understand that response and act accordingly – a problem that is larger, and less understood than the problem of simply detecting the emotional responses/expression in the first place.

There are many modern research systems that can be used to exemplify this concept in a classroom setting, including systems that detect emotion in speech (Polzin, 1998; Yu 2001), in facial expressions and gestures (Gunes 2005), in bodily cues (Strauss, 2005), and in text (Pang, 2002; Turney 2003). While there is a wealth of examples of projects in this space, I typically introduce the notion of detecting emotion by presenting my own work on detecting emotion in text (Owsley, 2006; Sood, 2007).

Expressing Emotion

Emotional expression within computer systems is typically focused on applications involving speech and/or facial expressions/gestures. Again, a plethora of work exists in this space, all of which would engage students in a classroom setting, including systems that attempt to automate gestures and expressions for an avatar (Breazeal, 1998; Breazeal, 2000), and those that enhance emotional expression through computer generated speech (Cahn, 1990). To introduce students to this concept, my work in the latter serves as a good example of machines that express emotion. This work is presented in a digital theater installation called *Buzz* (Sood, 2007), a system that has now moved online (www.buzz.com). In building a team of digital actors for *Buzz*, I wanted to empower them to convey emotion in their computer-generated voices. Standard text-to-speech engines are rather flat, emotionless; they wouldn't make for a compelling performance. I chose to augment an off-the-shelf text-to-speech engine with a layer of emotional expression (Sood, 2007). The end result is a system that actual conveys emotion (consistent with the content of what they are saying) in its voice.

Embodiment

Finally, the embodiment of a system facilitates a more personal connection between machine and user. People often attribute other human characteristics to a system when it perceives it as somewhat human/animal looking. This not only results in emotional connections, but it makes users more forgiving when the system makes a mistake. Many online systems in e-commerce, tutoring and training applications have recently begun embodiment as a way to engage/connect with users.

While there are many example systems in this space ranging from robotic seals to game based avatars, I find the most compelling example to be Kismet (Breazeal, 1998; Breazeal 2000), a robot created in the humanoid robotics group at MIT. The reason Kismet is such as great example is that it is an embodied system that detects, manages and expresses emotion in a social interaction with a human. While this is not the only system in this space, I think it is a compelling example to introduce students to the state of the art in all aspects of emotional intelligence (Sood, 2008).

Conclusion

The topics, exercises and references that I've provided are intended to guide students through an exploration of artificial intelligence within an introductory cognitive science class. Turing and Searle's work gives students a historical understanding of debate around the question "Can machines think?" Recent work in the realm of emotional intelligence can be a compelling way to address the question "Can machines feel?"

References

Bower, G. H. (1981). Mood and memory. Psychology Today, June, 60-69.

M. M. Bradley and P. J. Lang. *Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings*. Technical Report C-1, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

Breazeal, C. (2000), "Sociable Machines: Expressive Social Exchange Between Humans and Robots". Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT.

Breazeal(Ferrell), C. and Velasquez, J. (1998), "Toward Teaching a Robot `Infant' using Emotive Communication Acts". In Proceedings of 1998 Simulation of Adaptive Behavior, workshop on Socially Situated Intelligence, Zurich Switzerland. 25-40.

Cahn, J. E. The Generation of *Affect in Synthesized Speech*. Journal of the American Voice I/O Society, 1990.

Damasio A.R.. *Descartes' Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam, New York: 1994.

Ekman, P. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, New York, NY: 2003.

ELIZA, http://www-ai.ijs.si/eliza/eliza.html, 2008.

Robert M. French, "The Chinese Room: Just Say 'No!'," Proceedings of the 22nd Annual Cognitive Science Society Conference (2000), pp. 657–662.

Gardner, H. *Multiple Intelligences: New Horizons*. Basic Books, New York: 1993.

Goleman, D. *Emotional Intelligence: Why It Can Matter More Than IQ*. Bantam, New York: 1997.

Gunes, H., and Piccardi, M. *Fusing face and body gesture for* machine recognition of emotions. IEEE ROMAN Robot and Human Interactive Communication, 2005.

Jabberwacky, http://www.jabberwacky.com/, 2008.

JFLAP, http://www.jflap.org/, 2008.

Kihlstrom, J., and Cantor, N. Social Intelligence. in R.J. Sternberg (Ed.), Handbook of intelligence, 2nd ed. (pp. 359-379). Cambridge, U.K.: Cambridge University Press, 2000.

LeDoux, J.E.. *Emotion: Clues from the Brain*. Annual Review of Psychology, January 1995, Vol. 46, Pages 209-235.

Liu, H., Lieberman, H., and Selker, T. *A model of textual affect sensing using real-world* knowledge. In Proceedings of the 8th international conference on Intelligent user interfaces, 2003.

Mayer, J.D. & Salovey, P. 1993. The intelligence of emotional intelligence. Intelligence, 17, 433-442.

Mehrabian, A. Pleasure-arousal-dominance*: A general framework for describing and measuring individual differences in Temperament*. Current Psychology, Vol. 14, No. 4. (21 December 1996), pp. 261-292.

Norman, D. Emotional Design: Why we love (or hate) everyday things. Basic Books, New York: 2004.

Owsley, Sara, Sood, S., Hammond, K. *Domain Specific Affective Classification of Documents*. AAAI Spring Symposia Computational Approaches to Analyzing Weblogs 2006.

B. Pang, L. Lee, and S. Vaithyanathan. *Thumbs up? sentiment classification using machine learning techniques*. In Proceedings of EMNLP, pages 79–86, 2002.
Picard, R.W.. *Affective Computing*. MIT Press, Cambridge, 1997.

Polzin, T., and Waibel, A.. *Detecting Emotions in Speech*. In Proceedings of the CMC, 1998.

Russell, S.J., Norvig, P. Artificial Intelligence: A Modern Approach. Prentice Hall, New Jersey: 1995.

Salovey, P. & Mayer, J.D. 1990. Emotional intelligence. Imagination, Cognition, and Personality, 9, 185-211.

Searle, J.R. "Minds, Brains, and Programs," Behavioral and Brain Sciences, Vol. 3, (1980), 417–424.

Sood, S., Owsley, S., Hammond, K., and Birnbaum, L. *Reasoning Through Search: A Novel Approach to Sentiment Classification*. Northwestern University Tech Report Number NWU-EECS-07-05, 2007.

Sood, Sara Owsley. Compelling Computation: Strategies for Mining the Interesting. PhD Thesis, 2007.

Sood, Sara Owsley. *Emotional Computation in Artificial Intelligence Education*. Submitted to AAAI AI Education Workshop, 2008.

Strauss, M., Reynolds, C., Hughes, S., Park, K., McDarby, G. and Picard, R.W. "The HandWave Bluetooth Skin Conductance Sensor," The 1st International Conference on Affective Computing and Intelligent Interaction, October 22-24, 2005, Beijing, China.

Thorndike, E.L. 1920. Intelligence and its use. Harper's Magazine, 140, 227-235.

Turing, A.M. "Computing Machinery and Intelligence," Mind, New Series, Vol. 59, No. 236 (Oct., 1950), pp. 433–460.

Turney, P.D., and Littman, M.L.. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4):315–346, 2003.

Weizenbaum, J. *ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine*. Communications of the ACM Volume 9, Number 1 (January 1966): 36-35.

Yu, F., Chang, E., Xu, Y.Q., Shum H.Y.. *Emotion Detection from Speech to Enrich Multimedia Content*. In Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing table of contents, p550 – 557, 2001.