Artificial Intelligence: Past, Present, and Future

An independent literature survey conducted by Emily Wasylenko, guided by Dr. Dellinger

***Abstract*—This literature survey provides a basic philosophical foundation of AI, a summary of several branches of the field, and a brief speculation on the field’s future trajectory. The information given includes the topics of qualia, artificial moral agents, conceptual spaces, computational narrative, machine learning, nonmonotonic logic, and fuzzy logic.**

Introduction

Artificial intelligence is a broad field with a wide range of applications. Ever since the term was first coined in 1956, AI has generated a large amount of interest, and an equally large amount of research. Unfortunately, the sheer volume of research in this field makes it difficult to perform a comprehensive survey during a one-semester study. This paper attempts to achieve breadth over depth by dividing the entire field into five general sections, each briefly introducing an important topic of AI. These five sections are as follows: (1) philosophical issues with the field, (2) a proposed model for how the mind might work, (3) artificial moral agents, (4) blending and computational narrative, and (5) belief systems.

I. What is a mind?

The field of AI is as philosophical as it is technological—to build a mind, one must first understand what the mind is and how it functions. Many models have been proposed of the human brain; however, modeling the human mind is a much more abstract task. This section presents some problems and possible solutions introduced by two of the field’s leading scholars.

*A. Problems in the mind*

In his book Mind, John Searle [1] attempts to explain the workings of consciousness. First, Searle introduces the main problems with the philosophy of the mind. Second, he addresses arguments for and against materialism (a popular view that claims everything in existence is physical) and property dualism (the view that mental properties and physical properties inhabit nonoverlapping realms). Third, he gives his thoughts on consciousness itself, its structure and relation to neurobiology, along with how it relates to the mind-body problem. Finally, he presents the idea of intentionality. The remainder of the book attempts to solve these problems.

Here, I will summarize seven main problems with the philosophy of the mind, according to Searle. Firstly, the mind-body problem concerns the mysterious connection between the mental realm and the physical realm. Why and how are there causal relations between the two realms? How can anything physical (such as stepping on a Lego) produce an effect on something mental (the feeling of pain and a worsened mood)? How can events in the mind, such as a desire to raise one’s hand, produce a physical result?

It is difficult to understand one’s own consciousness, but at least an individual has direct access to his own thoughts and sense of consciousness. How can he be sure that other people he meets have minds of their own? This is the second problem addressed by Searle, the problem of other minds. It is possible to infer that others have consciousnesses by observing one’s own consciousness, but scientifically proving that others have consciousness is a different matter.

Just as one can be skeptical about other people’s consciousness, one can go a step further and doubt the existence of everything else. This skepticism about the external world is connected to another problem, the analysis of perception. When I look up at something, and my brain perceives that it is a bird, how can I know that there is an actual bird in front of me? The analysis of my perception could be faulty, as could the perception itself. How can I be sure I am not simply hallucinating everything in the external world? How can I trust my perceptions?

Everyone has made up their mind about something, that is, everyone has experienced deciding. At least, everyone thinks they have. However, it is difficult to prove that this sensation of free will is not an illusion. The fifth problem introduced by Searle, the problem of free will, is an amalgam of the mind-body problem and the problem of skepticism about the external world.

So far, Searle has stated problems of doubting other minds, of doubting the external world, and of doubting personal perceptions of the external world. In the sixth problem, he introduces the problem of the self and personal identity. How can you know that all your memories are yours? How can you know, after waking up, that you are still the same person that you were before going to sleep? If you wake up one day in a completely different body, are you still you? What makes you, you?

The final relevant problem is that of intentionality. Intentionality is the ability of a mind to refer to external things. When you think about an apple, you have produced a mental state in your mind that is about the apple. If thoughts in the mind are viewed as events, how is it possible for these events to refer to things beyond the brain? Why and how do intentional states have the specific contents they do have?

According to Searle, these problems stem from the two leading views of existence: property dualism and materialism. These problems should not be viewed as ultimately unsolvable simply because we do not have the right perspective on them. Consciousness should not be looked at as some mystical other-dimensional thing, nor should it be completely disregarded. To bridge the gap between the two viewpoints, Searle proposes a new view of reality: biological naturalism.

Biological naturalism states that conscious states are real and are caused entirely by neuronal processes in the brain. Conscious states arise as emergent properties of combined neuron activity. A singular neuron on its own is not conscious, but a large group of them acting together can create a conscious state. Since these conscious states are real features of the real world, caused by neurons in the real world, the changing of such states will have causal effects on actions or events in the real world.

Although biological naturalism cannot fully explain every problem of consciousness (for example, it cannot prove to you that I am conscious), it is important to learn from what Searle is doing. Problems of this nature may appear mysterious and beyond our grasp, but this does not mean they should not be questioned and prodded. A solution may eventually be discovered. Perhaps we have not yet developed the tools or the correct way of asking the question. Our current inability to grasp how neurons give rise to feelings does not doom the study of AI to futility.

*B. Qualia*

David Chalmers put it best, there are the easy problems of consciousness, and there is the hard problem. [2] The easy problems include programmable tasks such as problem solving, pattern recognition, and memory. The hard problem is sense experience, an ability to ask and answer, “What does it feel like?” Why does the color red feel different than the color orange? Why does the smell of cinnamon or the sound of nails on a chalkboard make us feel a certain way? Such sensations are called qualia. They are impossible to describe without using other qualia, and it is currently impossible to model or code their functions.

Rather than explaining qualia, Dennett chooses to explain it away in Quining Qualia. [3] Note that Dennett is not trying to explain away reality itself, nor is he trying to explain away the existence of conscious experience. He is trying to prove that nobody can be sure they have an accurate notion of what is happening inside their mind. According to Dennett, the burden of proof is on the experiencer. He uses fifteen “intuition pumps” to eliminate the intuitive problem of qualia, a substance that is ineffable, intrinsic, private, and directly or immediately apprehensible in consciousness. This paper will address three main ones.

The first intuition pump is the tastes of various foods. Why does orange juice taste sweet until you eat a doughnut? Why do some people enjoy the taste of cilantro, and others do not? Why do some people’s tastes change over time? All these questions prompt doubt that we can truly trust our sense of taste. The same questions apply to other qualia as well—why does a room feel warm until you jump in and out of a hot tub? How can we trust our sense experiences to be infallible when they change so often? A possible answer could be that our experiences, though they do change, change consistently. For example, orange juice always tastes sweet unless it is tasted immediately after drinking something sweeter. We can always trust this to happen. Additionally, the sensation of the “orangeness” of the juice is still there, whether we deem it sweet or sour. Further intuition pumps will be needed.

Suppose two people are looking at a mug of coffee. The mug is white. The coffee is dark brown. Do both people perceive the colors of this object the same way? This is the question posed by the “Inverted spectrum” pump. Everyone has been taught that objects are to be called a certain color, so our names for colors will be the same regardless of how we mentally perceive them. At present, there is no way to test this; but in the meantime, we should be cautious of how trustworthy our sense experiences really are.

Suppose we did have some kind of technology that showed us how other people saw the world, an apparatus called the “Brainstorm machine.” Upon connecting your mind to the experiences of another mind via the Brainstorm machine, you see every color in the world inverted. Is this proof that your qualia are different than another’s? Perhaps not; such an issue could be fixed by calibrating the Brainstorm machine to match up with your expected experience. At any rate, the public, physical objects you view, the sky, the ground, the birds, are the same ones looked at as those of another mind. Perhaps that other mind is biased by its own past experiences and slightly different neurobiological arrangement.

Dennett concludes that what seem to be private and indescribable qualia are simply illusions detected by our “property sensors.” The properties we detect are all public—when two people look at a glass of orange juice and experience a feeling, that feeling would not be possible without the public object (the orange juice) and its various public properties, all of which could be described in terms of other things in the external world. The only privacy here appears to be the idiosyncratic nature of our separate brains. This is an unsatisfactory conclusion. In the end, although Dennett would disagree, the hard problem of consciousness still appears to be unanswered. If these sense experiences are nothing more than illusions, why do they feel like anything at all?

II. Minsky and society of mind

“The society of mind” is a term coined by Marvin Minsky [4] to describe a theory of what comprises the human mind. Like the body is made up of many cells, the mind is made up of many abstract entities. These entities are strongly connected to their neighboring components, but just as the connections between most people grow weaker with increasing distance, so do the connections in the mind with increasing difference. Difference refers to the disparity between functions performed by these components. For example, there is little cross-communication between the entities controlling your foot movements and the entities in charge of hearing. Still, even unrelated members of the mind must influence each other somehow, or else we would not have this sense of a unified conscious state.

How do the members in the society of mind influence each other? The mind’s processes continually change themselves, but it is difficult to see how this is accomplished. Minsky notes that we know little about our own minds. We think without thinking about it, and it is difficult to understand how thinking really works if our thinking about thinking is not done without thinking about it. We understand so little about thinking that we use concrete metaphors like “conceiving,” “digesting,” or “chewing over,” to more fully explain the actions taking place inside our heads.

The ultimate goal of strong AI is to develop a computer that is sentient, conscious, and self-aware. Such a goal would be difficult to achieve not only because we cannot tell how our consciousness works, but because we do not even know what it means to be conscious. Minsky offers a broad description of consciousness as an “ability to sense what is happening in our minds, right at the present time.” If consciousness means knowing the events inside one’s own mind, then we have very weak consciousness indeed. Trying to see the events going on inside our minds, Minsky says, is like trying to photograph a subject that is moving too quickly. Not only that, this subject must, by nature, move whenever the camera’s eye is focused on it. Whenever we think about our thoughts, the thoughts change shape before our minds’ eye. This makes experiments regarding our own minds very difficult.

Although there are many roadblocks in the way of creating conscious machines, Minsky believes it might be possible to create machines that are even more conscious than we are, if consciousness is defined as in the last paragraph. This is due to short term memory. Our capacity for short-term memory is very low, and such memories easily morph and change if we think about them too closely just after they have arisen. Thus, it is difficult to understand the motions of our mind. If we could create a machine with greatly increased capacity for static, short-term memory, along with the ability to distance itself and analyze its thoughts and memory, then we could theoretically create a machine with greater consciousness than humanity itself.

III. Artificial moral agents

Although consciousness itself is still beyond our comprehension, a core ability of conscious minds, moral action, is not. As AIs become more and more ubiquitous, taking on roles from healthcare to the military, it makes sense that some form of morality should be implemented into their code. What would be the best strategy of implementation? The slightest moral failure could result in a life or death situation. In their paper Prolegomena to any future artificial moral agent, [5] Allen, Varner, and Zinser explain the risks and benefits of, and what it means to be, an artificial moral agent (AMA).

The first question they address is one of ethical framework. Which moral standard should AMAs follow? Some ethical theorists think a utilitarian standard would work best, defining good or evil based off the consequences or net happiness caused by an action. Others disagree, claiming the ends do not always justify the means, and some actions are morally reprehensible regardless of their good outcome. Instead, an action can only be morally good if it is consistent with Kant’s idea of the categorical imperative. The authors of the paper define the categorical imperative as a negative test, forbidding an agent to act on a maxim unless it is willing that its action become a universal standard. This gives an agent some capacity for failure, which raises the question: Can an AI be a true moral agent only if it holds some degree of autonomy?

To determine whether an AMA is sufficiently moral, the authors introduce the concept of a Moral Turing Test (MTT). If the test questions given to both human and computer focused only on morality, the ease of determining human from computer would measure how strong of a moral agent the computer was. Testing like this could help construct the perfect balance of utilitarianism and categorical imperative when designing an AMA. The main difficulty with the MTT is that the computer must not appear overly morally perfect or depraved in comparison to the human. Should AMAs be programmed to always make the moral choice, or should they be programmed more like humans, where they sometimes make an immoral choice? In some ways, it makes sense to think of AMAs differently than humans: human agents are impotent, and their moral flaws do not always lead to life-threatening situations. However, the nature of AMAs makes it vital that they behave in a morally good way. Perhaps it is a good thing for an AMA to consistently fail the MTT, if failing means moral perfection.

After listing some possible moral frameworks, the authors attempt to choose the best implementation. Programming explicit moral rules and principles would be a complex and difficult task. Modeling human behavior would be a much more desirable path. Three approaches to this modeling are given: virtue approaches, associative learning, and evolutionary/sociobiological approaches.

Virtue approaches value character over deeds since. The reasoning behind this is that good character precedes and produces good deeds. There are some problems with this approach. What should be done in situations when virtues conflict? When faced with a decision between honesty and compassion, which should an AMA choose? Additionally, it is difficult to define and quantify every existing virtue. Virtue approaches may be just as difficult to implement as a detailed list of programmed rules.

Associative learning likens the moral training of an AMA to that of a human child. The AMA is trained, via feedback, to behave in a morally correct way. Implementation frameworks such as neural nets would work well with this sort of approach. If would be difficult to determine the correct types of feedback for such training; however, if the training process were broken down into small steps, it might be feasible to determine the correct feedback through trial and error.

Evolutionary/sociobiological approaches combines sociobiology and game theory to produce correct decisions. This approach values reason above everything else, and trains AMAs to act in their best interest, where best interest is defined as giving the appearance of morality.

There is still much work that needs to be done in this area of AI, and it is unlikely that the perfect AMA will emerge until its human designers reach a consensus on which moral framework is the best.

IV. Blending and computational narrative

Conceptual blending is the action of taking several (possibly unrelated) inputs and combining them to make a new element, called a blended space. The blended space still bears characteristics of the inputs that went into creating it; however, it also has emergent properties that set it apart from other spaces, including the inputs that went into creating it. When implemented in an AI, some surprising results are achieved, as well as an improved ability for computational narrative.

A blended space is a difficult concept to imagine without a concrete example. Here is one provided by When given the two inputs “house” and “boat,” the obvious blended space would be a houseboat. This object shares some attributes of the two original inputs, and yet has a new meaning of its own. This is not the only blended space available, however. Another result could be the boat-house, a house inhabited by boats [6].

The five characteristics of optimal blends are as follows: integration (the blended space’s scenario should be well-integrated), web (the inputs must be closely related to the blend), unpacking (given the blend, it should be easy to figure out what the inputs and network of connections look like), topology (elements in the blend should act the same way as their counterparts in the input), and good reason (every element must have a meaningful place in the blend).

Computational narrative is the process by which an AI, given a few inputs and a framework of variable detail, constructs a story, poem, or song. As it turns out, conceptual blending is a very useful tool when it comes to computational narrative. In his article about the GRIOT System, Harrel [7] describes a computer program capable of creating poetry that nearly sounds as though it was written by a human. GRIOT accomplishes this with an algorithm called ALLOY, a conceptual blending algorithm implemented by Harrel.

With the help of ALLOY and human inputs, GRIOT is able to create poems such as the one italicized below. This poem, entitled “The Girl with Skin of Haints and Seraphs,” was created from inputs “Europe,” “Africa,” “girls,” “whiteness,” “devils,” and “seraphs.”

*her arrival onto this earth was marked when first-born and charcoal-girl transforms to impoverished-elder or charcoal-woman*

*she worked raising snow-queen original-lady children of her own*

*the young lady would prevail*

*a caress across her skin scares up demon black*

*her failure was ignoring her wings and original-lady nature*

*and she felt glad*

*as she grew older she saw entitlement defiance wrinkles upon her face*

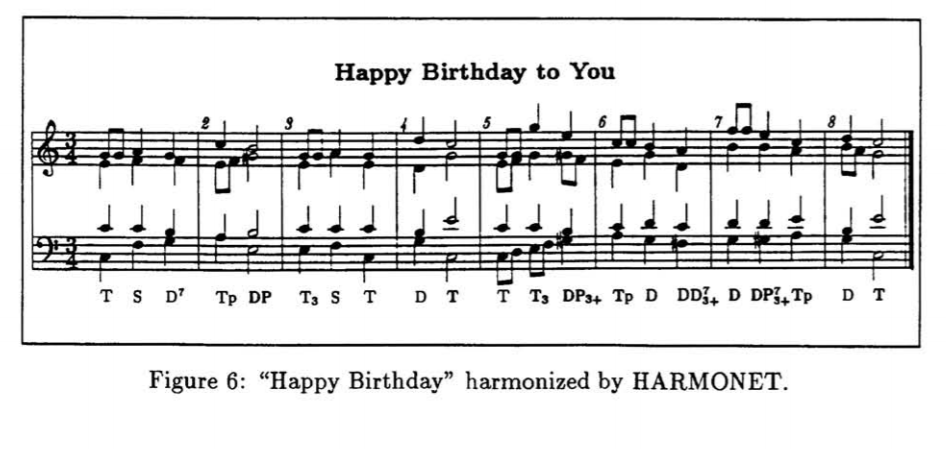
*ebony-wood-like brimstone defines fetish bedrock, the sign that let her know she finally*

*really alive*

As previously mentioned, computational narrative is not only useful for generating the written word. In a paper by Hild, Feulner, and Menzel, [8] they present a computer program called HARMONET that makes interesting four-part harmonies. HARMONET is a neural net that was taught the patterns of Bach chorales via error backpropagation. Error backpropagation is a technique that measures the correctness of the output and adjusts the neurons in the net accordingly.

HARMONET’s goal is to approximate the function that maps chorale melodies to their respective harmonies. To do this, the net is given a chord skeleton (the notes of the original piece with the eighth and sixteenth notes taken out). Inversion and dissonance are treated as fields of the chords. Each chorale was divided into smaller parts so that HARMONET could more easily learn them. Unsurprisingly, this fractured the overall consistency of the songs. To overcome this setback, HARMONET takes into consideration the window preceding the current window of focus, and the way previous chords affect subsequent ones.

After HARMONET was trained, it was given some melodies that were not in the training set. After creating the harmonies, HARMONET’s output was presented before an audience of music professionals. The audience declared the neural net to have the talent of an improvising organist. Below is a sample of HARMONET’s output when given the tune “Happy Birthday.”



V. Belief systems and logic

An obvious aspect of AI is logical thinking and belief: should AI be made to reason more like computers, or more like humans? What should be the structure of their belief system? In his paper, Doyle [9] describes two approaches to belief revision, foundations and coherence, and introduces the “reason maintenance system.” The coherence approach states that a static belief is held just so long as it logically coheres with every other belief it holds. The foundations approach states that beliefs are derived from justifications or evidence, and beliefs change as new evidence is presented.

Doyle brings up three operations that can be made on belief states through the coherence approach. Expansion and revision both deal with adding beliefs to a system. Revision differs from expansion in that expansion can result in consistent or inconsistent belief states, but revision must always produce a consistent state. The third operation that can be performed on belief states is contraction, which removes a belief from the system to create a new belief state.

The coherence approach treats every belief as independent of the other beliefs in the reason maintenance system. The foundations approach, on the other hand, splits all beliefs into two categories. The first category consists of beliefs derived from, or justified by, other beliefs. The foundations approach requires that every belief in this category be supported by a set of basic, foundational beliefs. Such foundational beliefs, which do not need to be justified by other beliefs, comprise the second category.

A reason maintenance system (RMS) organizes an agent’s set of beliefs by recording changes and results of changes to the RMS over time. The RMS has two elements: nodes and reasons. Nodes represent information such as beliefs, desires, rules, and procedures. Reasons represent inferences—rules that dictate the way nodes within the RMS influence each other. Every reason has an inlist and an outlist comprised of nodes. The nodes in the inlist support their reason, while those in the outlist oppose the reason. If the outlist is empty, the corresponding reason is monotonic. Otherwise, the reason is nonmonotonic. The latter type of reasoning can lead to some interesting results, explored in the field of nonmonotonic logic.

Nonmonotonic logic is a form of reasoning often utilized by the human mind. In nonmonotonic logic, the conclusions drawn depend on the order in which evidence was presented to the reasoner. This leads to some conclusions becoming invalidated after more evidence is presented. For example, assume Richard Nixon is a Quaker and a conservative. If you are first told that Nixon is a Quaker, you assume he is a pacifist. However, if you are told Nixon is republican, you would assume he is pro-military. Humans seem to place stronger weight on the evidence presented earliest. Incorporating this idea into an AI’s reasoning toolbox could produce more humanlike behavior.

In his paper, Dung [10] develops a way to mathematically implement nonmonotonic reasoning. He makes note of the aforementioned “Nixon Diamond,” as well as the stable marriage problem. The stable marriage problem considers two sets *M* and *W* of *n* men and *n* women respectively, where each member of one set has expressed their preference of marrying each member of the other set. A stable solution to this problem would be a scenario in which everyone is married to their first preference. Formally, this is the one-to-one correspondence *S*: *M*🡪*W* such that there exists no pair (*m*, *w*) in the set of *M* x *W* such that *m* prefers *w* to *S*(*m*) and *w* prefers *m* to *S*-1(*w*).

Often in real life, such a situation does not come to stable resolution. Thus, Dung concludes that even if a logic circuit comes up with an unstable result, this does not mean it has an error in it. Sometimes there is no perfect solution. Using this sort of logic in AI would be a good way to simulate real-life decision making.

Future work

If I could extend this independent study, I would explore the philosophical side of AI in greater depth. Additionally, I would like to read about temporal logic, chaos theory, and neural nets, especially the areas where neural nets behave insufficiently. Another interesting study might be the combination of neural nets, neurology, and some aspects of psychology.

Conclusion

The literature surveyed this semester are by no means a comprehensive picture of the field of AI. However, some broad conclusions can be drawn. A lot of work has been done using conceptual blending and computational narrative—the two sources presented in this paper are only the tip of the iceberg. The large volume and diversity of existing sources combining conceptual blending and neural nets indicates that this is one of the most well-developed areas of AI. On the other hand, the least developed portion of the field is the nontechnical, nonphysical matter of mind philosophy. Searle, Dennett, and Minsky present interesting allegories of how the mind appears to work. Some of their ideas are cohesive, others are contradictory. However, the philosophers’ conclusion is the same: they do not fully understand what a consciousness is. Certainly, there is room for improvement here, but this is the case with any field. Whether the capacity to create artificial life is a realistic dream or not, the field of AI has proved to be an interesting and useful one.

[1] J. Searle, “Mind, A Brief Introduction,” Oxford University Press, 2004.

[2] D. Chalmers, “Facing up to the problems of consciousness,” *Journal of Consciousness Studies,* 1995. [Online] Available: <http://cogprints.org/316/1/consciousness.html> [Accessed May 2018]

[3] D. Dennett, “Quining Qualia,” *Consciousness in Modern Science*, Oxford University Press, 1998. [Online] Available: <http://cogprints.org/254/1/quinqual.htm> [Accessed May 2018]

[4] M. Minsky, “The Society of Mind,” *The Personalist Forum*, vol. 3, no. 1, University of Illinois Press, 1987.

[5] C. Allen, G. Varner, J. Zinser, “Prolegomena to any future artificial moral agent,” *Journal of Experimental & Theoretical Artificial Intelligence*, July 2000.

[6] M. Eppe, E. Maclean, R. Confalonieri, O. Kutz, M. Schorlemmer, E. Plaza, K. Kühnberger, “A computational framework for conceptual blending,” *Artificial Intelligence*, March 2018. [Online] Available: <https://www.sciencedirect.com/science/article/pii/S000437021730142X> [Accessed April 2018]

[7] D. Harrell, “Shades of Computational Evocation and Meaning: The GRIOT System and Improvisational Poetry Generation,” Copenhagen, Demark, December 2005. [Online] Available: <http://groups.csail.mit.edu/icelab/sites/default/files/pdf/Harrell-DAC2005.pdf> [Accessed April 2018]

[8] H. Hild, J. Feulner, W. Menzel, “HARMONET: A Neural Net for Harmonizing Chorales in the Style of J. S. Bach,” *Advances in neural information processing systems,* 1992. [Online] Available: http://papers.nips.cc/paper/576-harmonet-a-neural-net-for-harmonizing-chorales-in-the-style-of-j-s-bach.pdf [Accessed April 2018]

[9] J. Doyle, “Reason Maintenance and Belief Revision: Foundations vs. Coherence Theories,” 1992. [Online] Available: https://www.researchgate.net/publication/2658662\_Reason\_Maintenance\_and\_Belief\_Revision\_-\_Foundations\_vs\_Coherence\_Theories [Accessed May 2018]

[10] P. M. Dung, “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games,” *Artificial intelligence,* 1995. [Online] Available: <https://ac.els-cdn.com/000437029400041X/1-s2.0-000437029400041X-main.pdf?_tid=9c6662d7-2bf7-4b62-8452-a12692daaefb&acdnat=1525903094_0796e445b9be6ea7a5449d99a8919d4b> [Accessed May 2018]