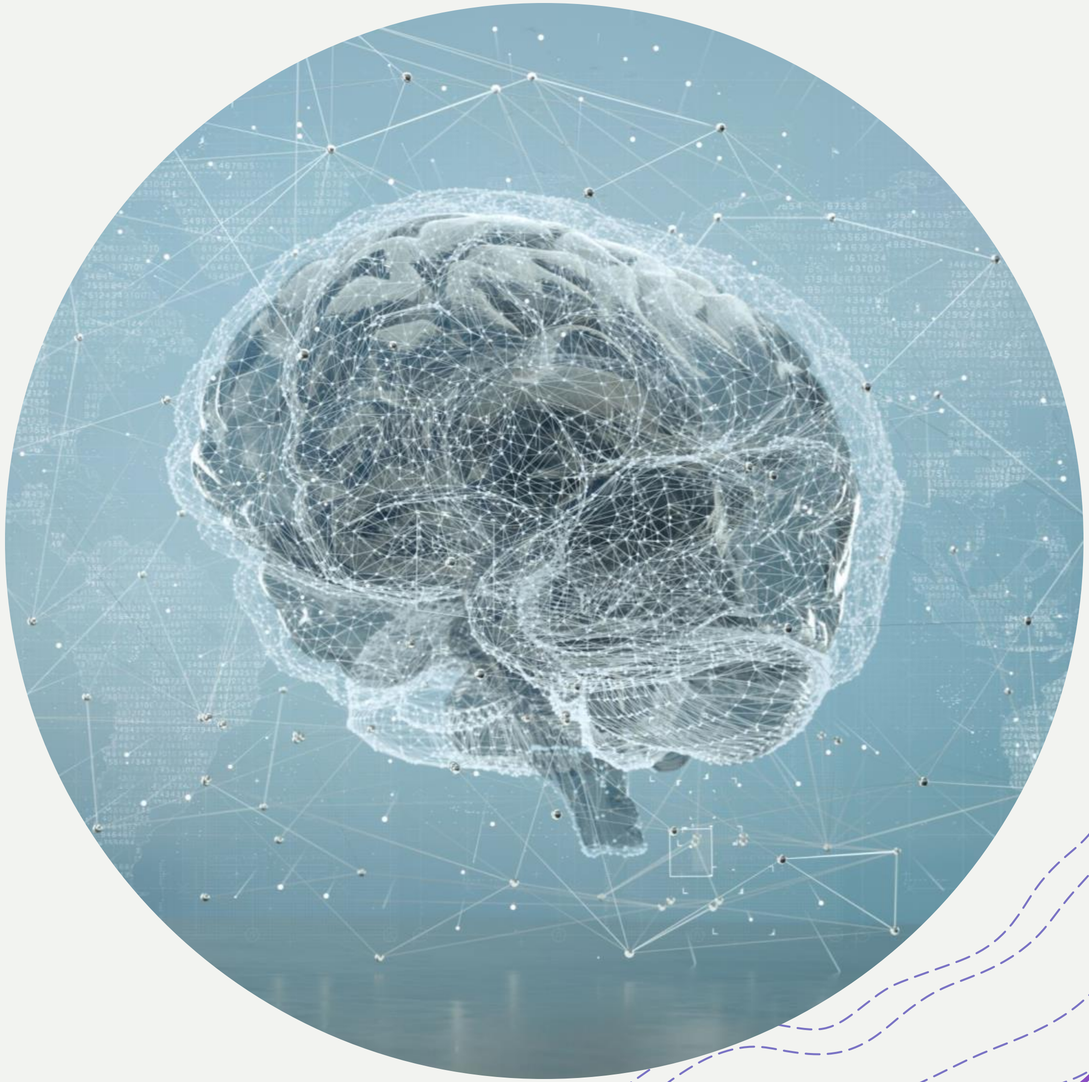
Intelligent Minds and Machines

PSYC 3043

Week 6: Language / Review / Project



office hours this week + Monday

- Thursday, 4.20 pm-5.30 pm
- Friday, 9 am-11 am
- Monday, 3 pm - 4.45 pm

today’s agenda

- language review!

what is the purpose of language?

- communicate
  - your needs (survival)
  - nuances / gestures/emotion
- to create bonds / listen to other’s needs
  - ask for help
  - caution, warn, etc.
  - cultural identity / affiliations
- to think (internal dialogue)
- entertainment / humor / sarcasm / storytelling

desiderata

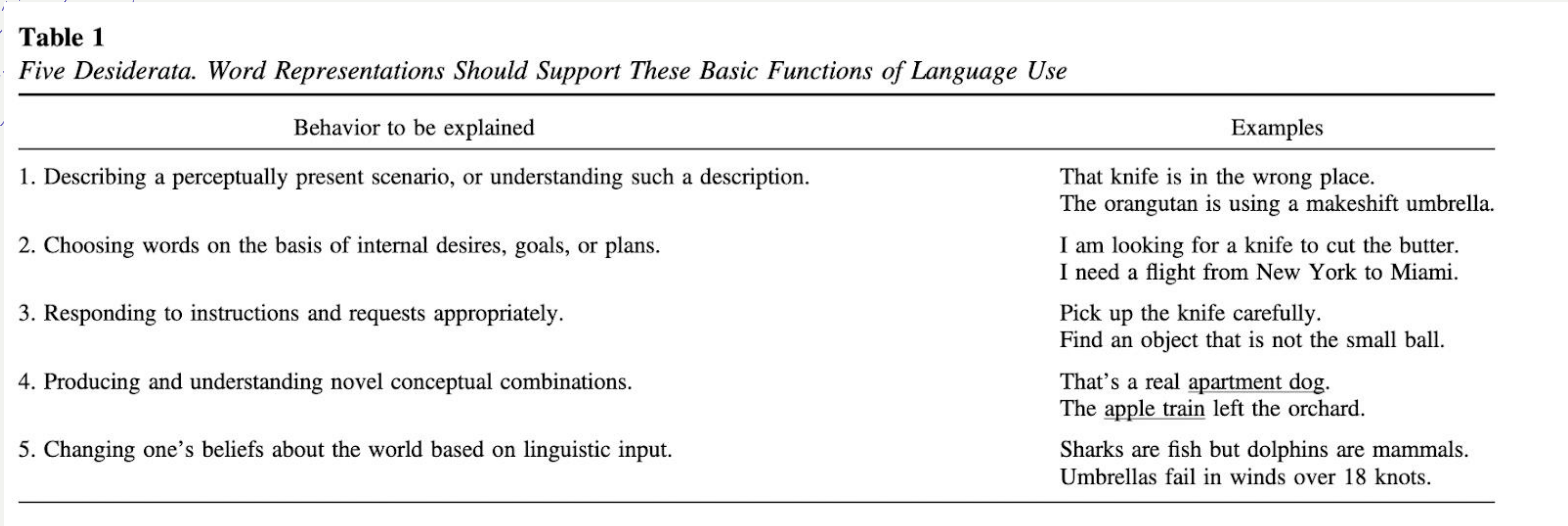


Table 1

Five Desiderata. Word Representations Should Support These Basic Functions of Language Use

Behavior to be explained	Examples
1. Describing a perceptually present scenario, or understanding such a description.	That knife is in the wrong place. The orangutan is using a makeshift umbrella.
2. Choosing words on the basis of internal desires, goals, or plans.	I am looking for a knife to cut the butter. I need a flight from New York to Miami.
3. Responding to instructions and requests appropriately.	Pick up the knife carefully. Find an object that is not the small ball.
4. Producing and understanding novel conceptual combinations.	That's a real apartment dog. The apple train left the orchard.
5. Changing one's beliefs about the world based on linguistic input.	Sharks are fish but dolphins are mammals. Umbrellas fail in winds over 18 knots.

why?

- why do we want to mimic human language in a model?
  - self-discovery (we learn about ourselves)
  - existential (we want something that we’re familiar with)
    - what we know
  - trust
  - humanizing / anthropomorphizing / bonding
  - efficiency / synthesize information / utility
    - iterating / testing / self-exploration
  - communication with humans
    - better understand / follow instructions

Why NOT?

- why do we NOT want to mimic human language in a model?
  - reduce human input / devalue human expertise
  - losing social connection
  - misinformation / biases / expert fallacy
  - uncanny valley / creepy / sentience threat
  - climate / environmental costs
  - malicious uses (phishing / scams / deepfakes)

why?

- why do we care about comparisons between:
  - animals and humans
    - understanding ourselves (classical conditioning / associative learning)
    - abilities that animals are better at humans
    - decenter humans / undermine “humans are superior”
    - morality / ethical concerns
  - animals and machines
    - replicate simpler animals
    - consciousness continuum
  - humans and machines

explain/define the following

+ take some time to jot down an explanation, then we’ll come talk about it together

explain/define the following

- word representations

explain/define the following

- how does a language model learn what different words mean?

explain/define the following

- large language model

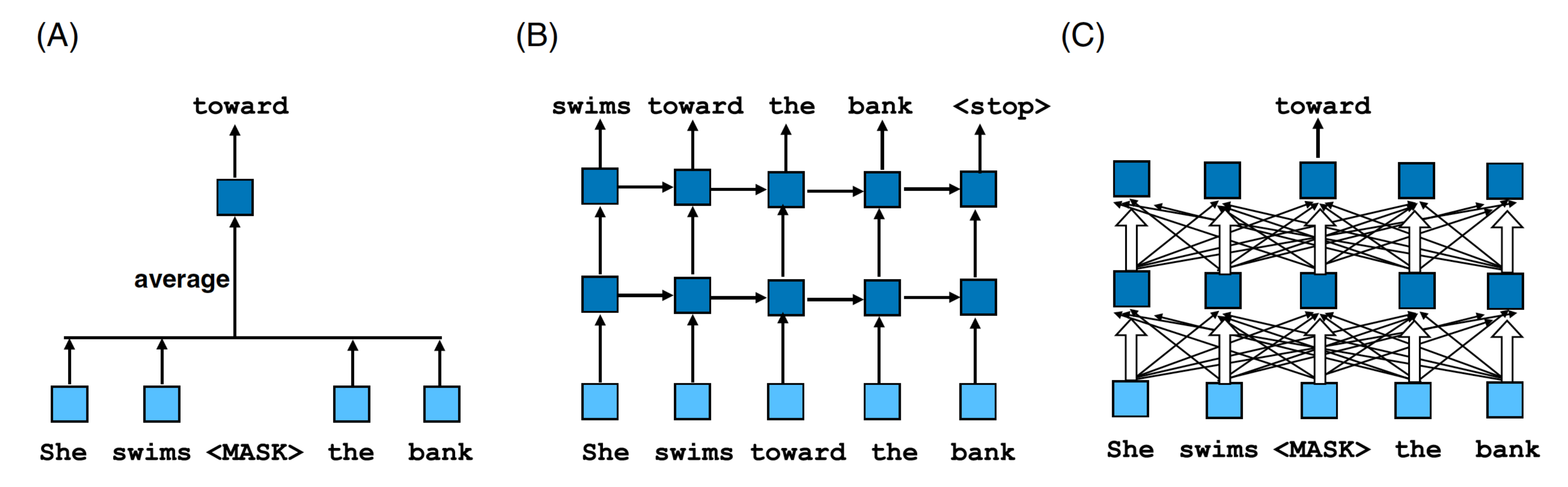
explain/define the following

- how does we test whether a model has learned language?
- what would convince you?

types of language models

- “count”-based: count how words co-occur, condense this information into meaningful representations
- “prediction”-based: are trained to predict something, and to do well on this task, they develop better representations
  - “transformer” architecture / “attention”: improves predictions by using MORE context words (before and after)
  - primary mechanism behind current large language models

types of language models



word2vec
no concept of word order,
averages all words' representations
within a window

RNN
predicts each upcoming word
using an indirect connection
to previous words

BERT/GPT-2
all words contribute to different
degrees in generating a target
word's representation

Searle (1999) Chinese room argument / also Harnad (1990)

Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.

mistakes

- Jennifer: how do the kinds of mistakes vary between humans and machines?
- Rachel: why does this sort of flaw persist

example

Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says “chocolate” and not “popcorn.” Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. Sam cannot read. Sam looks at the label.

Fill in the blank with one word: She believes that the bag is full of

____

implications

- Haley: Why is this important? Is there a more superior way to learn overall? What can humans learn from these machines?
- Ocean: How is generative AI is impacting different sectors like education, advertising, law, etc.? What are these implications and how can we best address and tackle these issues at places like Bowdoin and in higher education?
- Ocean: How are black box issues in AI currently being addressed?

making it human

- Emely: With the way we are trying to get AI to be more human, what is the necessary level of biological fidelity we want AI to have?
- May: Can a machine (i.e., ChatGPT) embody a persona? When I ask ChatGPT human-like questions, it can never respond. Why don't programmers just make ChatGPT a "person"?
- Rachel: I wonder how much of our understanding of these "emergent features" stems from our actual real-world interactions with these objects. (dirty bowls)

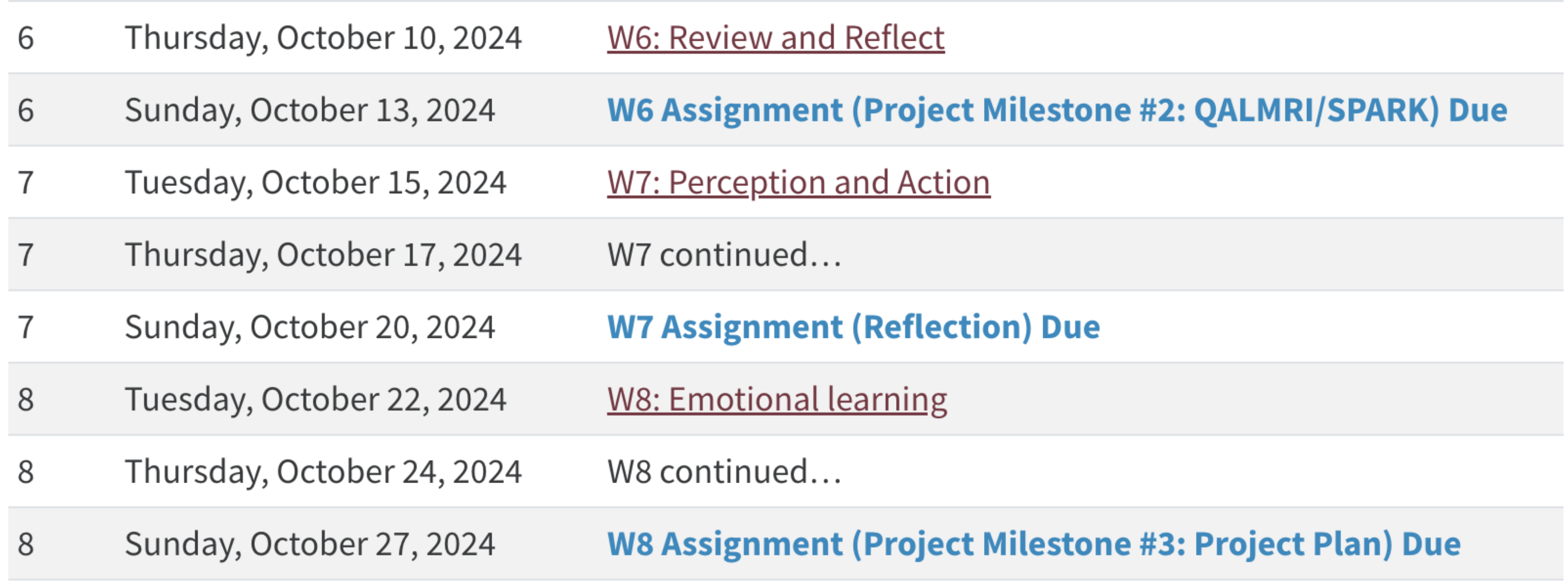
The Octopus Test, Bender and Koller (2020)

Imagine that person A and B are independently stranded on two deserted islands, but they can communicate with each other via an underwater cable and often send text messages in english to each other. Without either person A or B's knowledge another entity O (a very clever octopus) who cannot speak english but has a very advanced knowledge of statistics and pattern matching. After some very long time, O decides to cut the wire so that they can speak directly to each person. The question is, could O have learned enough from the form (the text messages) so that neither person knows that anything has changed?

going forward

- “embodiment”: going beyond “amodal” symbols that are not grounded in the real world

project discussion

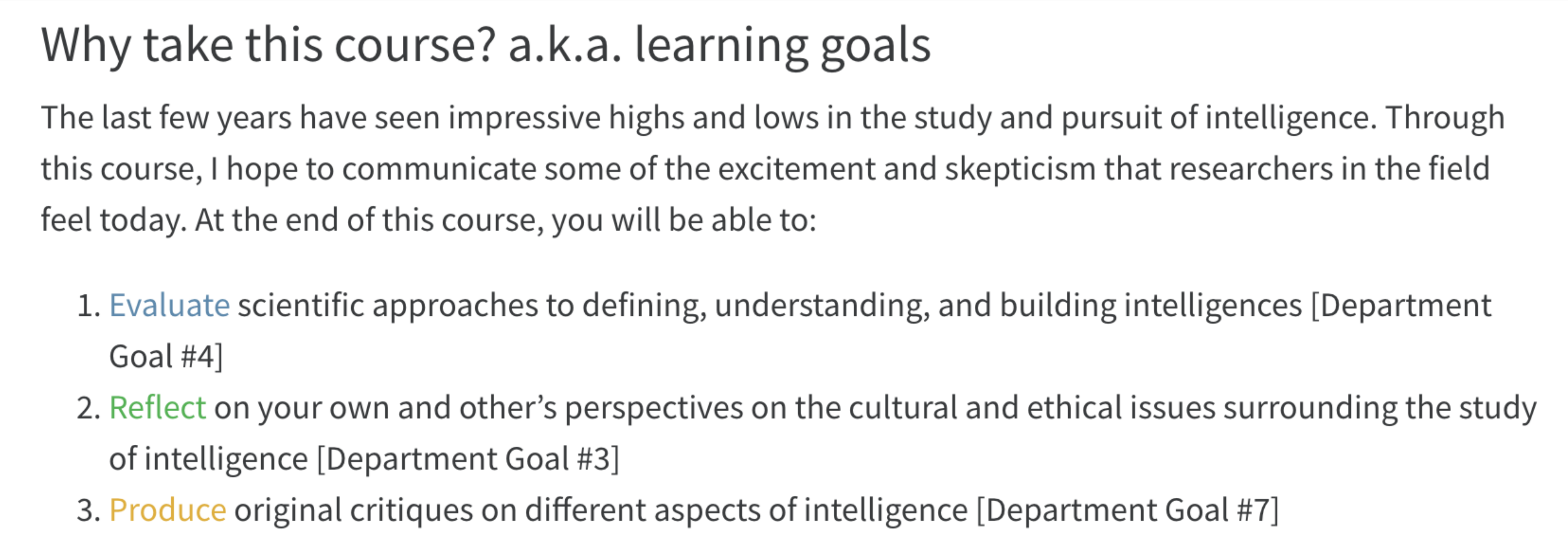


6	Thursday, October 10, 2024	W6: Review and Reflect
6	Sunday, October 13, 2024	**W6 Assignment (Project Milestone #2: QALMRI/SPARK) Due**
7	Tuesday, October 15, 2024	W7: Perception and Action
7	Thursday, October 17, 2024	W7 continued…
7	Sunday, October 20, 2024	**W7 Assignment (Reflection) Due**
8	Tuesday, October 22, 2024	W8: Emotional learning
8	Thursday, October 24, 2024	W8 continued…
8	Sunday, October 27, 2024	**W8 Assignment (Project Milestone #3: Project Plan) Due**

project discussion

- this week:
  - find 1 review article and submit SPARK
  - find 5 empirical articles and submit QALMRIs
- discuss:
  - your current interests/topics
  - Qs about finding articles
  - what formats are you considering?

project discussion



Why take this course? a.k.a. learning goals

The last few years have seen impressive highs and lows in the study and pursuit of intelligence. Through this course, I hope to communicate some of the excitement and skepticism that researchers in the field feel today. At the end of this course, you will be able to:

1. Evaluate scientific approaches to defining, understanding, and building intelligences [Department Goal #4]
2. Reflect on your own and other's perspectives on the cultural and ethical issues surrounding the study of intelligence [Department Goal #3]
3. Produce original critiques on different aspects of intelligence [Department Goal #7]

project discussion

- what would you like to take away from this course?
- what formats do we want to allow that meet these learning goals?

- Advertisement
- Brochure
- Game
- Interactive exhibit
- Letter
- Media analysis
- Monologue
- Op-ed
- Piece of fiction
- Podcast
- Recipe for nostalgia
- Research proposal
- Scrapbook
- Video compilation/montage
- Visual artwork
- Webpage
- Zine