Course Syllabus-Relying on hallucinations: The linguistics behind human AI interactions

Language of Instruction: English
Professor: Martina Wiltschko
Professor's Contact and Office Hours:

Office: Campus Poblenou, Roc Boronat 138, 52.629
Martina.wiltschko@upf.edu
For office hours please e-mail me for an appointment
During the week of the class, I will be available for consultation after the class (13:00-14:00)

Course Contact Hours: 15 hours
Recommended Credit: 2 ECTS credits
Weeks: 1
Course Prerequisites: None
Language Requirements: Recommended level in the European Framework B2 (or equivalent: Cambridge Certificate if the teaching language is English, DELE or 3 semesters in the case of Spanish)
Course structure: Seminar
Course classification: Introductory or Advanced

Course Description:

ChatGPT frequently produces false information: output that appears plausible but is not factual. This is known as 'hallucinations'. The reason behind this is the fact that large language models (LLM) are trained to predict strings of words (rather than being a repository of 'facts'). Crucially, an AI does not "know" about the truthfulness of its output. Nevertheless, AI-tools are increasingly used to provide "information" in professional and private settings. Why are we inclined to rely on this non-reliable source?

In this course we explore this question from a linguistic angle. We compare the logic and architecture behind LLMs (which underlie AI-tools) with the logic and architecture behind human cognition (including the capacity for language). At the root of our "trust" in AI-tools is the apparent flawless language output, which can lead to anthropomorphization, which in turn leads users to expect that it follows the same conversational principles as humans do.

In this course, we explore several aspects of human language that contribute to our inclination to take AI generated output at face value:

i) the fact that meaning in language is based on truth-conditions;
ii) the fact that humans mark uncertainty with linguistic means that are conspicuously absent in AT-generated text;
iii) the fact that human communication is based on the cooperative principle (according to which we assume that our interlocutors are reliable).

As limiting cases, we will compare the virtual hallucinations of AI-tools with pathological hallucinations in Schizophrenia, as well as language that does not rely on truth (poetry).

Learning Objectives:

At the end of the course, the student

1. will have developed an understanding of the linguistic factors that contribute to the perception of truth in AI generated text even when false information is presented.
2. will be able to identify the difference between AI generated text and linguistic interaction between humans.
3. will have improved critical thinking skills in in the use of AI and in general.

Course Workload:

The course will consist of short lectures, and in-class activities that will lead to developing research questions, hypotheses, and testing these hypotheses via self-observation and the design of online surveys. Students should be prepared to read approx. 100 pages in this course, as well as describe and analyse their observations and collected data.

Methods of Instruction:

I will use a mixed and blended teaching methodology consisting of

i) **Short lectures** on some of the core concepts students will have to understand (including *linguistic knowledge in humans; the linguistics behind large language models; truth conditions; language beyond truth: poetry and pathological halucinations*; *expressing certainty vs. marking uncertainty in human communication; linguistic cooperation that guides interpretation).* These lectures will be based on the flipped class-room style where students will have read papers and watched a relevant video *before* class)

ii) 5 min **videos** covering the essence of the core concepts which will discussed in the lectures. (The videos will be produced should the course be accepted.) A sample video can be found here.

iii) Hands-on **workshop**-type activities comparing interaction with ChatGPT and human-human interaction (based on self-observation and watching selected video-clips of conversations). For example, relying on native speaker judgments we will compare the "well-formedness" of an interaction intended for a human with one intended for an AI-tool.

Classes will be held in collaboration with some of my PhD students who will assist in facilitating discussions and hands on workshop activities.

Method of Assessment:

Class participation [20%]

Online quizzes [20%]

These have to be completed before each class and are based on understanding of videos and readings

Final project [60%]

Case-studies of an instance of a GPT hallucination using the variables that influence our perception of these hallucinations as "truth" introduced in the course. The results of the project can be presented as a paper (traditional academic style; 12 pages) or as a short video (outreach style; 5 minutes).

Absence Policy:

Attending class is mandatory and will be monitored daily by professors. The impact of absences on the final grade is as follows:

Absences	**Penalization**
Up to one (1) absence	No penalization
Two (2) absences	1 point subtracted from final grade (on a 10 point scale).
Three (3) absences	The student receives an INCOMPLETE for the course

The BISS attendance policy does not make a distinction between justified and unjustified absences. All absences—whether due to common short-term illnesses or personal reasons—are counted toward the total amount and cannot be excused. Therefore, students are responsible for managing all their absences.

Only in cases of longer absences—such as hospitalization, prolonged illness, traumatic events, or other exceptional situations—will absences be considered for exceptions with appropriate documentation. The Academic Director will review these cases on an individual basis.

Classroom Norms:

· No food or drink is permitted.
· There will be a ten-minute break during the class.
· Students must come to class fully prepared.

Course Contents:
Please, detail here the course topics distributed on a weekly or daily schedule. Example:

Day 1(July 21, 2025)
Syllabus, Assessments, readings
Introducing and exemplification of the problem: Hallucinations in AI

An overview of the **linguistic capacities** of humans vs. AI: human cognition vs. large language models.
Reading: Zanotti, G., *et al.* (2023).

Day 2 (July 22, 2025)
An introduction to **meaning** in human language (part 1)
truth-conditions and compositionality
Testing the limits: poetry, lies, and pathological hallucinations)
Readings: Munn, et al. (2023); Parnas et al. (2023)

Day 3 (July 23, 2025)
An introduction to **meaning** in human language (part 2)
The **expressive** power of human language:
How do we talk about (un)certainty?
Reading: Wiltschko (2022)

Day 4 (July 24, 2025)
An introduction to the use of language
Cooperation in linguistic interaction: humans vs. machines
Reading: Dombi et al. (2022)

Day 5 (July 25, 2025)
Conclusions
Presentations of project results
Lessons in **critical thinking** and information consumption

Required Readings:

Dombi, J.; Sydorenko, T.; Timpe-Laughlin, V. (2022). Common ground, cooperation, and recipient design in human-computer interactions, Journal of Pragmatics 193: 4-20, https://doi.org/10.1016/j.pragma.2022.03.001
Munn, L., Magee, L. & Arora, V. Truth machines: synthesizing veracity in AI language models. *AI & Soc* (2023). https://doi.org/10.1007/s00146-023-01756-4
Parnas, J.; Yttri, J.-E.; Urfer-Parnas, A. 2023. Phenomenology of auditory verbal hallucination in schizophrenia: An erroneous perception or

something else?, *Schizophrenia Research* (online first) https://doi.org/10.1016/j.schres.2023.03.045

Wiltschko, M., (2022) "Language is for thought and communication", *Glossa* 7(1). doi: https://doi.org/10.16995/glossa.5786

Zanotti, G., Petrolo, M., Chiffi, D. *et al.* (2023). Keep trusting! A plea for the notion of Trustworthy AI. *AI & Soc* https://doi.org/10.1007/s00146-023-01789-9

Recommended bibliography:

I will make additional optional readings available on a dedicated course-website.