Assignment 2
Essays

Introduction

In this assignment, you will write two brief essays, one on the importance of interpretability and one related to ethical aspects of AI in general (and conversational agents in particular).

2.1 Interpretability (10p, mandatory)

As mentioned in Chapter 1 in the compendium, interpretability is crucial in high-stakes decision-making and, by extension, also in high-stakes human-computer dialogue, for example in cases where a CA provides medical advice or assists in (semi-)automated driving. The importance of interpretability has increased dramatically with the public release of several black-box systems for dialogue (or, more generally, text generation), such as Meta Galactica (now defunct) and chatGPT. While those models are both impressive and interesting, the widespread (and sometimes uncriticial) adoption and use of black-box conversational systems is likely to pose a signficant danger, given the lack of common-sense exhibited by statistical language models.

It is important also to distinguish between *interpretable models* where interpretability is an integral part of the model and so called *explainable AI* which does include interpretability, but is often focused on trying to give post-hoc explanations of the output given by black box models. Here, our focus is on (the importance of) models that are inherently interpretable.

1. Read Cynthia Rudin's paper *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, in Nature Machine Intelligence, **1**, pp. 206-2015 (2019), which you can access (online) via Chalmers' library or Google scholar.

2. For the specific case of conversational AI, read our paper on key principles for safe and accountable conversational AI, available at `https://dl.acm.org/doi/pdf/10.1145/3507623.3507632`.

3. Using Google Scholar, download and read some (3-10, say) papers on interpretable AI, for *example* papers that cite Rudin's paper or our paper. Make a careful selection of high-quality papers.

4. Then watch the following presentation

   `https://www.youtube.com/watch?v=zsRKPxgHURQ&ab_channel=StochasticProgrammingSociety`

and make sure that you understand the difference between an interpretable model (on the one hand) and the concept of trying to *explain* a black box model (see also above).

Then write a short essay on the importance of interpretability, taking into account the material that you studied in the previous steps. The essay should be at least 4 pages long, and at most 6 pages long, with 11 pt font and standard A4 page size, and with at most 2 figures (not larger than roughly one third of a page each). Use the following four section headings in your essay: 1. Interpretable models vs. black-box models, 2. Interpretability vs. explainability, 3. Dangers of black models, 4. Conclusion. Write your essay as you would write a chapter in a book, and emphasize *clarity* in your writing. After reading your essay (with section headings as just described), the reader should know (a) what are interpretable models, (b) what are black box models, (c) the difference between inherently interpretable models and attempts at explaining black box models, (d) why interpretable models are important in high-stakes decision-making. You should also (e) give some specific examples of the dangers associated with using black box models in high-stakes decision-making. At the end of the report, include references to the papers selected in Step 3 above, as well as references to any other cited papers.

What do hand in You should hand in your report in PDF format, following also the other requirements listed above.

Evaluation The essays will be judged both on your level of insight (regarding the importance of interpretability) as well as the clarity and structure of the essay. If you need to resubmit the essay, a maximum of 6p will be given.

2.2 Ethics (10p, mandatory)

In addition to interpretability, the ethical aspects of AI in general, and CAs in particular, have started becoming very important, with many new AI systems (such as the above-mentioned Meta Galactica and chatGPT) being rolled out for use in many different aspects of human affairs.

Moreover, while there are many positive aspects of smart phones and similar devices, they make it possible for e.g. companies and governments to gather massive amounts of information. In many cases this might not be a problem and, indeed, many people are not particularly concerned about such issues. However, it might *become* a problem if the body that collects and holds the data has malevolent intent (or, perhaps, is targeted by other organizations, companies, or governments who might steal the data and use it for unethical purposes).

Moreover, the black box nature of many AI systems, along with their data-driven training approach may itself lead to various problems, as exemplified by Microsoft's Tay agent, as well as the more recent failures of large SLMs in certain situations. Other problems involve, say, the tendency of black box systems to exhibit various forms of biases (e.g. racial biases). Some additional examples are (i) the recent advent of so called deep fakes, whereby an AI system generates say, speech or even videos, which are then falsely attributed to some person or organization, and (ii) the possibility of swamping social networks with CAs that have malicious intent, e.g. spreading disinformation.

AI technology has the potential to offer many benefits, but researchers (and students) in this field also have an obligation to avoid unethical uses of AI systems (for example conversational agents) to avoid contributing to a dystopian, Big Brother-like future with, say, mass surveillance without the consent of those who are being monitored. You should now do the following:

1. First, read the articles listed under this assignment (2.2) on the Canvas course page.

2. Then, using Google Scholar, find an additional set of papers (3-5, say) on ethical issues related to the use of AI. (For example, *start with* the search phrases *ethics artificial intelligence*, *unethical uses of AI*, *malevolent uses of AI*, or similar phrases). The papers should then be included in the reference list in your essay, along with a specification of the number of citations (which *must* be obtained from Google Scholar) for each paper.

3. Next, as a case study, select a specific issue where an AI system (preferably, but not necessarily, a conversational agent) is either being used unethically or could potentially be used unethically (in the *near* future, i.e. no science fiction) for example by a commercial company (e.g. invasion of privacy), or a government (e.g. mass surveillance and its consequences). Then search for information on the selected specific topic. In this part, you can use a wide search (i.e. not only scientific papers available in Google scholar), but be careful (and critical) when evaluating the *source* of the information, preferably focusing on a topic for which there are *several independent sources* providing information (as one should of course always do).

Then write a short essay on ethics in AI taking into account the material that you studied in the previous steps. The essay should be at least 4 pages long, and at most 6 pages long, with 11 pt font and standard A4 page size, and with at most 2 figures (not larger than roughly one third of a page each). At the end of the essay, include a list of references, as described above. In case of scientific papers, make sure also to list their number of citations on Google Scholar. Use the following five section headings in your essay: 1. The importance of ethics in AI, 2. Examples of unethical uses of AI, 3. Case study: <TEXT>, where <TEXT> provides the title for your case study, 4. Preventing unethical uses of AI, 5. Conclusion. Write your essay as you would write a chapter in a book, and emphasize *clarity* in your writing.

When reading your essay, the reader should learn (a) why is it important to take ethical considerations into account when developing AI systems, (b) what are some of the ethical dilemmas in AI (and with CAs especially); (c) some examples of unethical uses of AI, and (4) the details of a specific (potential or actual) unethical use of an AI system (i.e. the topic that you studied in Step 2 above); (d) how unethical uses of AI can be prevented.

What do hand in You should hand in your report in PDF format, following also the other requirements listed above.

Evaluation The essays will be judged both on your level of insight (particlarly regarding the prevention of unethical uses of AI) as well as the clarity and structure of the essay. If you need to resubmit the essay, a maximum of 6p will be given.