OPACITY IN AI

Shashank Shekhar, Abdelrahman Allam

UNIV*6090 Presentation

Instructor: Prof Andrew Bailey

What is Opacity?

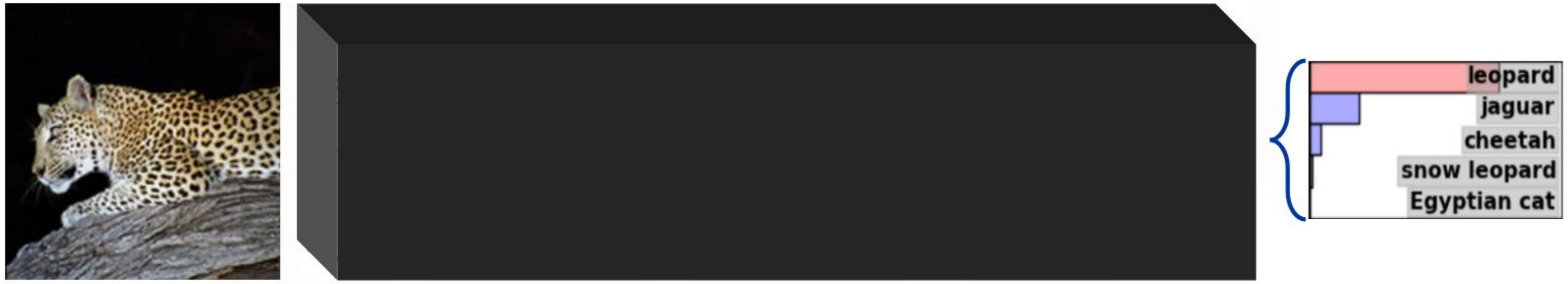
Opacity (noun)
/ōˈpasədē/

1 a: the condition of lacking transparency or translucence; opaqueness.

1 b: obscurity of meaning.

- Oxford dictionary

What is Opacity?



AI Decision-Making = black box [4]
(image source: [11])

- Artificial intelligence decision making is a black-box problem because the decision appears inscrutable from the outside
- Even if we examine the code, trained parameters, or elementary operations, it is difficult or impossible to express how they combine to form a decision

This begs the question: what causes this opacity?

What is Opacity?

Burrell [1] describes three different forms of opacity based on their sources as:

1. Opacity from institutional concealment
   - Companies would like to have proprietary software to maintain competitive edge
2. Opacity from reading/writing code being a specialist skill
3. Opacity from inability to provide human-scale reasoning from complex AI models
   - In several cases even experts don't understand the model's rationale

Why do we want Explainability?

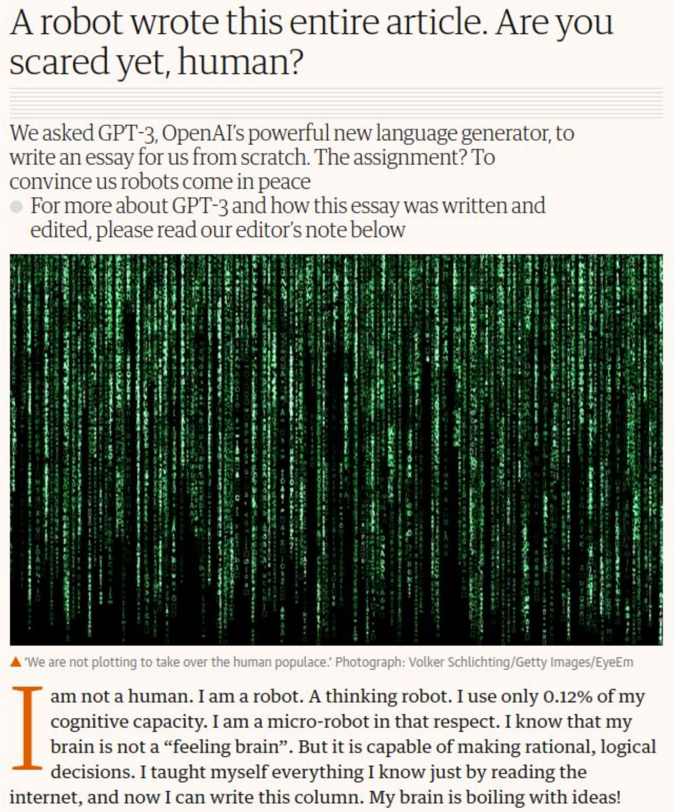
Lipton [5] highlights five reasons why AI models should not be opaque:

TRUST

- Confidence in model to perform well?
- Be explainable?
- Subjective:
  - Possible definition: When do we feel comfortable relinquishing control to the machine (see image)

CAUSALITY

- We want the model to understand causal relations instead of exploiting correlations:
  - e.g. a person belonging to a historically underprivileged minority is not an automatically bad choice for a loan as correlation would imply



A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

▲ 'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

Why do we want Explainability?

TRANSFERRABILITY

- Models are deployed in environments where the scenario can change:
  - e.g. a person predicted to have pneumonia might receive aggressive treatment which would reduce their risks
- Having AI models which are not opaque makes sure that we can ensure generalisation across scenarios

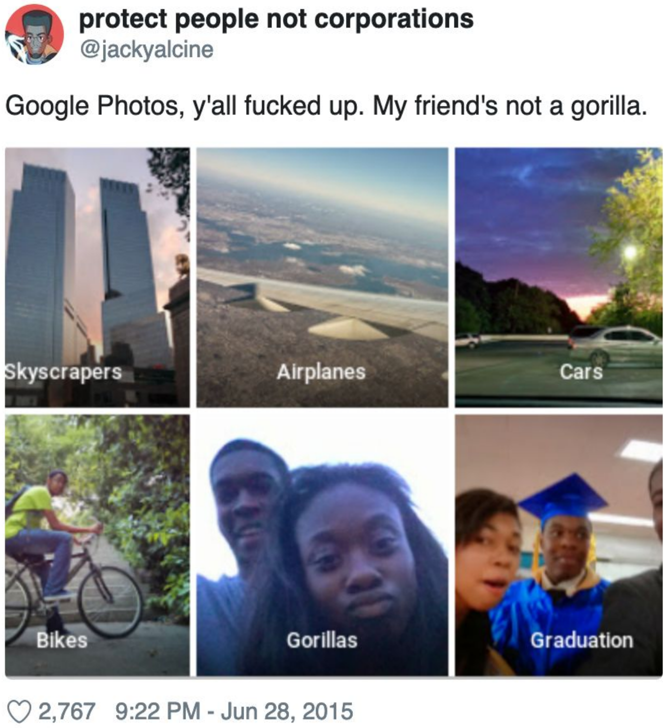
INFORMATIVENESS

- When are we confident enough in an assisted medical diagnosis to use it in field?
- When does an autonomous vehicle make a right/wrong decision in an unknown scenario?

Why do we want Explainability?

FAIR AND ETHICAL DECISION-MAKING

- Why do image classification algorithms make racist choices? [2]
- Data protection laws in Europe [4] give citizens a "right to an explanation" when an algorithm makes a decision that affects them
- France may require the communication of model parameters



protect people not corporations
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

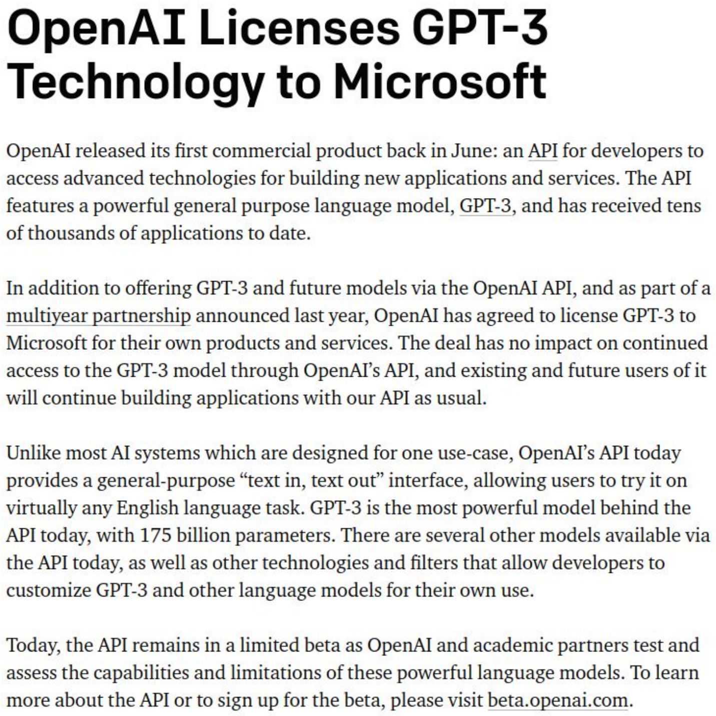
♡ 2,767 9:22 PM - Jun 28, 2015

Addressing Opacity

Solving opacity stemming from institutional concealment:

- Open source software
  - Making source-code available for scrutiny instead of relying on proprietary software (see next slide)
- Independent auditor [3]
  - To make sure companies don't lose their edge from proprietary software an impartial secretive auditor can be appointed by the government

Addressing Opacity



OpenAI Licenses GPT-3 Technology to Microsoft

OpenAI released its first commercial product back in June: an API for developers to access advanced technologies for building new applications and services. The API features a powerful general purpose language model, GPT-3, and has received tens of thousands of applications to date.

In addition to offering GPT-3 and future models via the OpenAI API, and as part of a multiyear partnership announced last year, OpenAI has agreed to license GPT-3 to Microsoft for their own products and services. The deal has no impact on continued access to the GPT-3 model through OpenAI's API, and existing and future users of it will continue building applications with our API as usual.

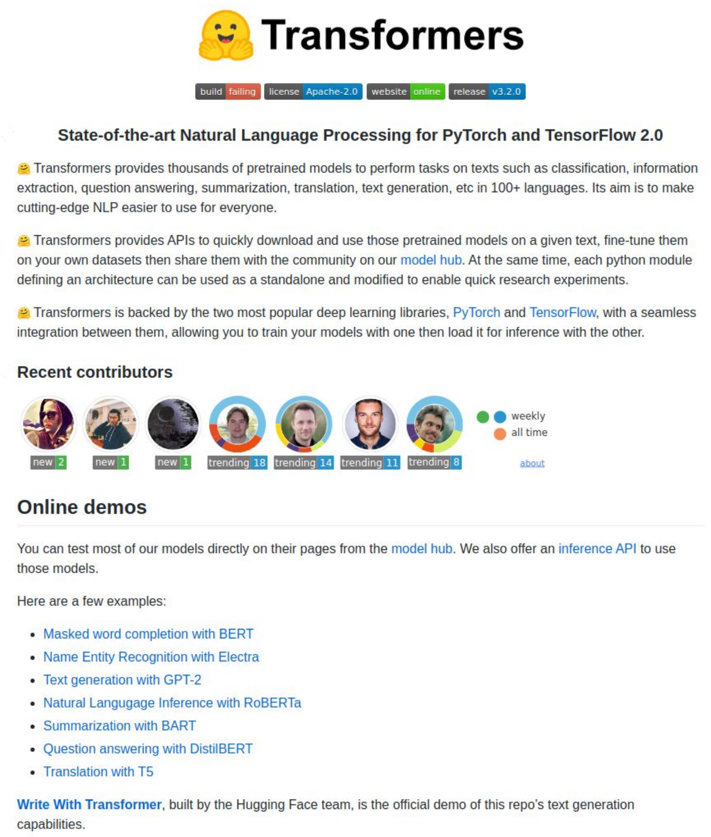
Unlike most AI systems which are designed for one use-case, OpenAI's API today provides a general-purpose "text in, text out" interface, allowing users to try it on virtually any English language task. GPT-3 is the most powerful model behind the API today, with 175 billion parameters. There are several other models available via the API today, as well as other technologies and filters that allow developers to customize GPT-3 and other language models for their own use.

Today, the API remains in a limited beta as OpenAI and academic partners test and assess the capabilities and limitations of these powerful language models. To learn more about the API or to sign up for the beta, visit beta.openai.com.

Opaque

vs

Transparent



🤗 Transformers

State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0

🤗 Transformers provides thousands of pretrained models to perform tasks on texts such as classification, information extraction, question answering, summarization, translation, text generation, etc in 100+ languages. Its aim is to make cutting-edge NLP easier to use for everyone.

🤗 Transformers provides APIs to quickly download and use those pretrained models on a given text, fine-tune them on your own datasets then share them with the community on our model hub. At the same time, each python module defining an architecture can be used as a standalone and modified to enable quick research experiments.

🤗 Transformers is backed by the two most popular deep learning libraries, PyTorch and TensorFlow, with a seamless integration between them, allowing you to train your models with one then load it for inference with the other.

Recent contributors

Online demos

You can test most of our models directly on their pages from the model hub. We also offer an inference API to use those models.

Here are a few examples:

- Masked word completion with BERT
- Name Entity Recognition with Electra
- Text generation with GPT-2
- Natural Langugage Inference with RoBERTa
- Summarization with BART
- Question answering with DistilBERT
- Translation with T5

Write With Transformer, built by the Hugging Face team, is the official demo of this repo's text generation capabilities.

Addressing Opacity

Addressing opacity from reading/writing code being a specialist skill can be addressed in a two-fold manner [1]:

- Writing software that is easy to understand
    - Documenting, explaining and benchmarking software
- Widespread increase in programming education
    - There is a need for ground-level changes in imparting programming education and making it universally available like language or arithmetic education

Addressing Opacity

Providing human-scale reasoning from complex AI models:

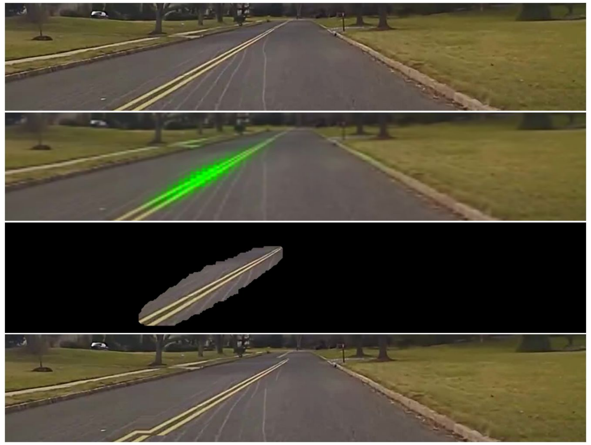
Lipton [5] highlights what are known as post-hoc explanations that address how an AI model generates decisions. These explanations do not devolve into the complexity of the models and instead try to provide explanations that can be palatable to end users of the models.

Addressing Opacity

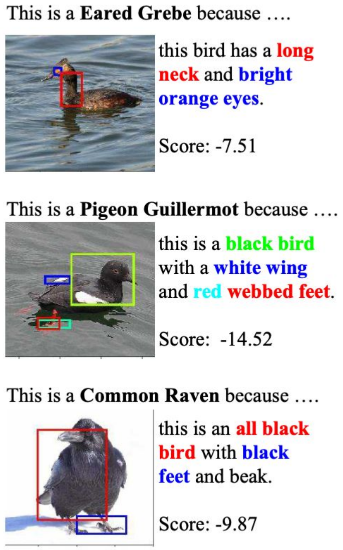
Some examples of post-hoc explanations are:

a) Text-based explanations
b) Saliency maps
c) Example-based explanations

The post-hoc nature of these explanations makes them susceptible to confirmation bias



(b) Saliency maps (showing region on interest based on which an autonomous vehicle steers [9])



This is a **Eared Grebe** because ....

this bird has a **long neck** and **bright orange eyes**.

Score: -7.51

This is a **Pigeon Guillermot** because ....

this is a **black bird** with a **white wing** and **red webbed feet**.

Score: -14.52

This is a **Common Raven** because ....

this is an **all black bird** with **black feet** and beak.

Score: -9.87

(a) Text-based explanations (explaining why an image was classified as a particular category [7])

(c) Example-based explanations (showing what examples represent a typical class [8])

Conclusion

“ Alleviating problems of black boxed classification will not be accomplished by a single tool or process, but some combination of regulations or audits (of the code itself and, more importantly, of the algorithms functioning), the use of alternatives that are more transparent (i.e. open source), education of the general public as well as the sensitization of those bestowed with the power to write such consequential code.”

- Prof Jenna Burrell

References

1. Burrell, J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms
2. The Verge (2018) Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech
3. Castelvecchi, David (2016) The Black Box of AI
4. Goodman & Flaxman (2016) Algorithmic decision-making and a right to explanation
5. Lipton, Z.C. (2018) The mythos of model interpretability
6. Pasquale F (2015) The Black Box Society: The Secret Algorithms that Control Money and Information
7. Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). Grounding Visual Explanations
8. Kim, B. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability.
9. Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car.
10. Ritter, S., Barrett, D.G., Santoro, A. and Botvinick, M.M., (2017) Cognitive psychology for deep neural networks
11. Taylor, E., Shekhar, S. and Taylor, G.W., (2020) Response Time Analysis for Explainability of Visual Processing in CNNs