DataLab 1: Implicit bias & A Designer's Critical Alphabet

Today, you will be introduced to the concept of implicit bias and made familiar with the Open Images dataset. First, we are going to conduct an experiment (i.e. Harvard's Implicit Bias Test). Next, we will perform an EDA on the images, and corresponding meta-data. Later, we will critically assess the dataset in terms of bias, and fairness by applying concepts from A Designer's Critical Alphabet.

Learning objectives

1. Define the term 'worldview'.
2. Explain how a worldview can be considered unique/personal, as well as shared.
3. Explain the difference between the terms 'implicit', and 'explicit' bias.
4. Assess the Open Images dataset in terms of bias, and fairness by applying concepts from A Designer's Critical Alphabet.

Table of contents

1. Q&A and standup: 1 hour
2. Introduction: 1 hours
3. Workshop: 4 hours
4. Additional material (optional): 1 hour
5. Reflection & Work/learning log: 1 hour

Questions or issues?

If you have any questions or issues regarding the course material, please first ask your peers or ask us in the Q&A in Datalab!

> Tip: Note down any important questions you might have!

Good luck!

---

1) Q&A and Standup

1a Ask questions regarding the independent study material.

1b Answer the following questions:

- What did you do yesterday?
- What will you do today?
- Are there any impediments in your way?

---

2) Introduction

In (social) psychology, bias is classified as either conscious ('explicit') or unconscious ('implicit'). Explicit bias occurs when people consciously devalue or hold negative views of a particular group. Implicit bias, on the other hand, refers to the brain's automatic, instantaneous association of stereotypes and attitudes with particular groups of people (Dovidio, Kawakami, and Gaertner, 2002). These implicit biases are often contrary to our personal values and are formed by the social environments around us (van Nunspeet et al., 2015).

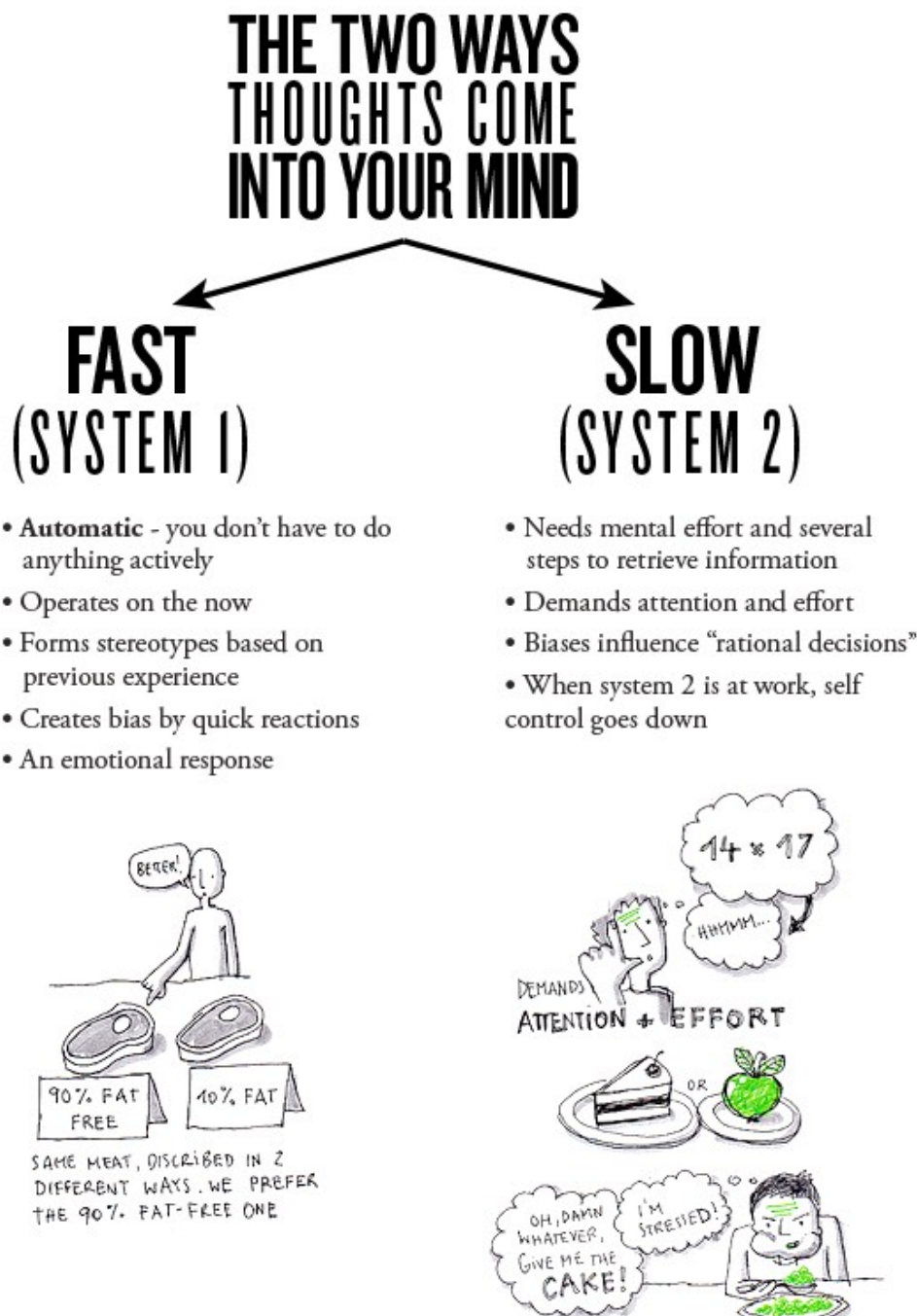Implicit biases are an example of what Kahneman and Tversky (1974) call 'System 1' thinking.

THE TWO WAYS THOUGHTS COME INTO YOUR MIND

FAST (SYSTEM 1)

- **Automatic** - you don't have to do anything actively
- Operates on the now
- Forms stereotypes based on previous experience
- Creates bias by quick reactions
- An emotional response

SLOW (SYSTEM 2)

- Needs mental effort and several steps to retrieve information
- Demands attention and effort
- Biases influence "rational decisions"
- When system 2 is at work, self control goes down

Figure 1. Systems of thinking (Kahneman, and Tversky, 1974).

This mode of thinking is fast, emotional, and unconscious. It requires little effort, but is prone to error.

Let's focus on a more concrete example of implicit bias: Ethnic profiling. A Dutch police officer may be committed to 'protect and serve' its community, and perceive itself as impartial, yet at the same time, they can

lead a stop and frisk police action that targets a specific ethnic group (e.g., Dutch citizens of Turkish descent). Consciously, the police officer is deeply loyal to protecting the community; however, their behavior is inconsistent with those beliefs because they are showing prejudice against certain community members.

Figure 2. Implicit bias.

The ability to hold these two truths (e.g., 'serve and protect the community', and 'unequal treatment of community members based on ethnicity') and not see them is what makes implicit biases so dangerous.

2.1 Implicit bias test

First, we need to uncover our implicit biases. To do so, we are going to conduct a little experiment:

2.1a Take the Implicit Bias test by Harvard University, which you can find, here.

2.1b Reflect upon your result of the Implicit Bias test:

- Are we all bigots?
- What is more dangerous implicit or explicit bias?
- etc.

In-class group exercise.

Unfortunately, we cannot cure ourselves from our implicit biases, but with self awareness we can address them, and hopefully make them less prominent.

2.2 Worldview

Now we have identified our implicit biases, we can look at the possible causes. Some scholars argue that many of the common prejudices of people brought up in 'The West' (partially) originate from having a Eurocentric worldview.

Worldview (definition):

> We can understand worldview as a mental model of reality. It contains a comprehensive framework of ideas and attitudes about the world, ourselves. Our worldview is shaped by our inherited characteristics (identity), lived experience, values, attitudes, habits and more. These elements of ourselves vary from one person to another. Therefore, even though some parts of a worldview are shared by many people in a community, other parts differ for individuals, so worldviews are both shared and unique (Source).

2.2a To refresh you memory, scan through the article Eurocentrism by Keena Hays, which was part of yesterday's DataLab preparation.

2.2b How does the Mercator map fit within the Eurocentric worldview? In your answer, refer to the concept of 'bias' and/or 'fairness'. Write your answer down.

Figure 3. Mercator world map projection.

2.2c Give at least three examples of how a worldview such as Eurocentrism can lead to bias in a dataset? Write your answer down.

3) Workshop: Open Images & A Designer's Critical Alphabet

In the workshop, we are going to investigate how the Eurocentric worldview (i.e., the driving force behind many of our implicit biases) affects famous benchmark image datasets such as Open Images.

To do so, we will use A Designer's Critical Alphabet, which is a tool that helps you to reflect on the design process (of a data science/AI project) by applying concepts of critical theory; a philosophy that, amongst others, involves challenging the prevailing view of society.

By critiquing the status quo, in this case the Eurocentric worldview, we allow for a space where group differences, and multi-cultural viewpoints can be acknowledged, and appreciated; a shared worldview that encompasses for many personal/unique worldviews.

To make it a bit more concrete, let's explore the Open Images dataset, and see how this dominant worldview influences it:

If you look at images with the class label `Wedding`, you will see that they often depict a traditional western wedding (e.g., heterosexual couple, white/ivory wedding dress, etc.). A classifier trained on the Open Images dataset will reflect this particular bias; images that deviate from the norm are more likely to be misclassified than images that adhere to the norm.

To mitigate this bias we need to make the data more varied (e.g., 'inclusive'). There are multiple ways to achieve this goal. For example, we could add additional images of weddings that do not adhere to traditional western norms to the dataset (e.g., LGBTQ+ couples).

First, we need to familiarise ourselves with the Open Images dataset:

3a Start exploring the Open Images dataset by performing an EDA on the images, and corresponding meta-data (e.g., images, and image-level labels).

To make your lives a bit easier, we created a Jupyter Notebook template for the Responsible AI Datalabs 😁. You can find the template, here.

When performing the EDA, try to answer the following questions:

- Which class label might be prone to bias?
- What types of bias can you identify in the dataset?
- At what stage of the CRISP-DM cycle is this particular type bias introduced?
- Whom or which group might be disproportionally affected by this particular form of bias.
- How can we mitigate this particular type of bias?

When you are finished with the EDA, form groups of three, and proceed with the following question:

3b Pick at least three cards from A Designer's Critical Alphabet that will help you, and your teammates to make the Open Images dataset more 'inclusive'. Elaborate on your choice. Notify the lecturer when finished.

- A Designer's Critical Alphabet

In-class group exercise.

> Note: Need further information or have questions? The lecturers will be available throughout the DataLab session.

---

4) Additional material (optional): Decoloniality

Decoloniality is a school of thought, which

> ... offers an option for thinking and doing beyond the dominant paradigms. it provides a critical analysis of modernity understood broadly as the western project of civilization, while it seeks to overcome the dominion of western epistemology and aesthetics and their embedded eurocentrism and anthropocentrism (Rolando Vázquez, 2021, xvii).

4a Watch the following video: Decolonising Knowledge: What is Decolonisation? by Rolando Vázquez.
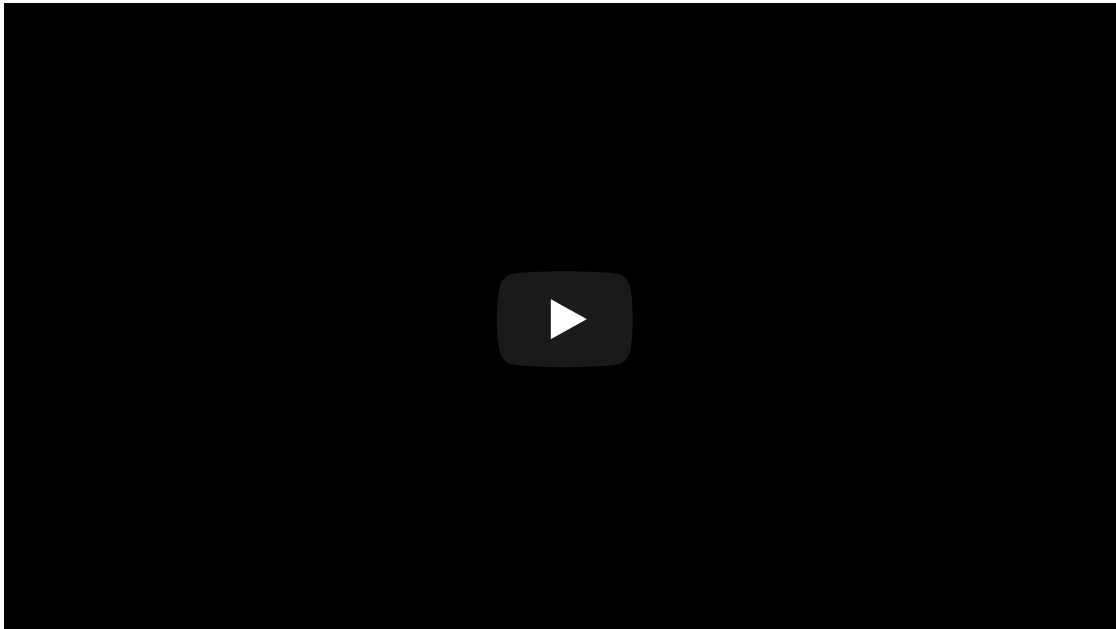
Video 1. Decolonising Knowledge: What is Decolonisation?

4b Define the term 'positionality', and provide a concrete example of the concept. Write your answer down.

4c What does Vázquez mean by 'truthful knowledge'? Explain your answer.

4d In the video, Vázquez puts a lot of emphasis on the importance of listening. Why does he do this?

Decolonial theory: Zone of being/non-being

As a result of colonization, the world was divided into two zones: one half represented the core (zone of being), while the other half represented the periphery (zone of non-being) (Fanon, et. al., p. 2). By dividing the world along these radical lines, a global hierarchy of superiority and inferiority was established:

Figure 5. Fanon's Conception of Racism.

> As a result of this divide everything south of the equator vanished, became non-existent: On the other side of the line, there is no real knowledge; there are beliefs, opinions, institutions, and subjective understanding, which at most, may become object or raw material for scientific inquiry. Thus, the visible line that separates science from its modern others is grounded on the abyssal invisible line that separates science philosophy, and theology, on one side, knowledges rendered incommensurable and

> incomprehensible for meeting neither the scientific methods of truth not their acknowledged contenders in the realm of philosophy and theology (Santos, 2014, p. 120)

European colonizers were seen as superior to the native population, which were often referred to as the 'other':

> We went from the sixteenth century characterization of 'people without writing' to the eighteenth and nineteenth century characterization of 'people without history', to the twentieth century characterization of 'people without development' and more recently, to the early twenty first century of 'people without democracy' (Source).

This systematic destruction, or 'war' on indigenous knowledges is called epistemicide. Furthermore, the colonizers secured their superior position through violent means: 'Beyond the equator are no sins' (Santos, 2014, p. 121), because copresence of the two sides was deemed impossible (Santos, 2014, p. 118).

What to learn more about decolonizing AI? See Mohamed's article Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, which '... explores the important role of critical science, and in particular of post-colonial and decolonial theories, in understanding and shaping the ongoing advances in artificial intelligence.'

---

5) Reflection & Work/learning log (16:00-17:00):

5a Fill in your work/learning log.

5b Choose, and provide an answer to at least of the following questions:

1. What surprised you today, and why?
2. What is the most important thing you learned today? Why do you think so?
3. What do you want to learn more about, and why?
4. When were you the most creative, and why do you think that is?
5. What made you curious today? How does learning feel different when you are curious?
6. When were you at your best today, and why?
7. (Assuming we were studying the same thing and you could decide and have access to anything), where would you start tomorrow? Why?
8. What can/should you do with what you know?

Resources

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659-684.

Santos, B. S. (2017). Epistemologies of the South: Justice against epistemicide. London: Routledge/Taylor&Francis.

Vazquez, R. (2021). Vistas of modernity: Decolonial aesthesis and the end of the contemporary. Amsterdam: Mondriaan Fund.