Background

v.0.0.1

1 Intro

We're going to be talking about topics that cut across a wide range of disciplines. That means we need to have some familiarity with a bunch of different concepts. Here's a very rough and very quick overview of some of them, just so we can all be on the same page.

2 Data

Basically everything we're going to discuss revolves around the use of data. So, what's data?

Very roughly, a piece of data (a datum) is a piece of information or a (purported) fact. More carefully, we'd probably want to say that data is information in the context of some use.[1] The temperature of the room you are in now is not data. It would be if we were adjusting the HVAC or running an experiment about how temperature affects reading ability.

2.1 Facts

Data represents facts. What are facts?

Facts represent a state-of-affairs which obtain. A state-of-affairs is the way part of the world is at a time. Take the sentence 'At 6 PM on 10 May 2019, Adam Swenson weighs 200 lbs'. This sentence is true because it says the world is a certain way and the world is actually that way. In other words, the sentence describes a possible state-of-affairs. Since that state-of-affairs obtains (it actually happens), the sentence describes a fact.

Consider three sentences:

There is a taco on Mars.
There is a taco in Adam's hand
The dog needs attention

All three are states of affairs. The first two are not facts. There are no tacos on Mars; I am (sadly) presently taco-less. The third is a fact. He keeps nudging me, bringing over toys, and otherwise acting cute in wanton disregard of my need to write this. Be right back.

It is important not to confuse the fact with what actually makes it true. The thing that makes the fact that I weigh 200lbs true is that if you add up all the masses of all the atoms comprising my body at the specified time within the Earth's gravitational field, you will get 200 lbs.

2.2 Representations of facts

Can you own a fact? What could that even mean? Facts are abstract things. The fact that I weigh 200lbs can be expressed with different sentences. For example, 'if you make a big pile of all the things which weigh 200lbs and look through it, you will find

[1] Some writers invert this formula and define information as data plus use; we needn't wade into this fight here.

Adam' (hopefully near the top). Thus no one can own a fact. Even if you owned me, you would not own the fact about my weight.

What you can own is a representation of a fact. Lets call that a datum. A datum can come in different forms (sentences, representations in a database). A datum (the singular form of 'data' which no one uses) is a (purported) fact.

If my weight is recorded in Hoolie's database and someone hacks into the database and erases the record, presumably Hoolie can sue her for destroying its property. I probably cannot. In some cases, the conditions under which the company acquires data gives the user rights to the data. The user agreement might say, for example, that if I cancel my account, Hoolie will delete all records related to me. In other cases, I may have no such right to the data Hoolie possesses.

3 Personal data

What's personal data? Data of a personal nature. All done. Moving on…

Okay, just kidding (although not by much). Here's a summary of what we are and aren't interested in from the Stanford Encyclopedia of Philosophy

> Personal information or data is information or data that is linked or can be linked to individual persons. Examples include date of birth, sexual preference, whereabouts, religion, but also the IP address of your computer or metadata pertaining to these kinds of information. Personal data can be contrasted with data that is considered sensitive, valuable or important for other reasons, such as secret recipes, financial data, or military intelligence. Data that is used to secure other information, such as passwords, are not considered here. Although such security measures may contribute to privacy, their protection is only instrumental to the protection of other information, and the quality of such security measures is therefore out of the scope of our considerations here. [2]

Thus personal data is a proper subset[3] of data. Namely, it is data about a natural person.More importantly for what we'll be talking about, an often-used definition is the one found in the European Union's Data Protection Directive, namely

[2] https://plato.stanford.edu/entries/it-privacy/

[3] A fancy way of saying all personal data is data, but not all data is personal data

> "Any information relating to an identified or identifiable natural person"[4]

Notice that this hinges on whether a piece of information or data can be explicitly related back to a person. [5]

3.1 Natural persons

Natural persons are contrasted with unnatural persons, for example, corporations; also, non-persons like rocks.

There may be borderline cases. If a robot turns out to be sufficiently like a human being that similar moral considerations should be extended to them, then perhaps the robot could become a natural person. For now, natural persons are limited to human beings.

3.2 Linkability

What it is to be 'about' a natural person? Is it enough that it represents a fact where a person is the subject. Or does the person have to be identifiable?

The ability to link a piece of data to a natural person is the decisive feature. There are 2 ways of making a link. Writers distinguish between referential and attributive uses.

Referential uses are made on basis of (possible) acquaintance relationship of the speaker with the object of their knowledge. For example, if someone says "the murderer of Tupac must be insane" while pointing at him in a courtroom, they are referring directly to a particular person. This is usually the sort of connection between a person and a piece of data which the law is concerned with.

Attributive uses, say something about a person without implying that we know anything about who they are. If I say "The murder of Tupac must be insane, whoever he is", there is no implication that I'm actually talking about someone I can pick out. If personal data is understood this way, most data will not be protected under current regulations.

4 Algorithms

If you ask a computer scientist or mathematician what an algorithm is, you'll get something like

[4] EU Data Protection Directive (95/46/EC) Article 2(a) [ToDo: check formatting]

[5] See Sax 30

> An ordered set of unambiguous steps that produces a result and terminates in a finite time.

That's more formal than we need. We'll just say that an algorithm is a stepwise computational procedure for doing something.

Outside of technical contexts, indeed, often in the articles we'll read, some writers use the term in narrower or more loaded ways. For example, some talk about algorithms as computational processes used to make decisions. Decision-making is a subset of the things we might do with algorithms.[6]

4.1 Adam's attendance algorithm

For a simple example, here's the attendance algorithm I follow at the beginning of every class:

- For each student, call out name. If answer, mark present
- When done, ask if anyone was missed
- If anyone answered 'yes', mark each answering student as present

If we were doing this with a python function, it might look something like (lines starting with '#' or enclosed between triple quotes """this is a comment""" are comments for humans to understand what's going on and not part of the program):

```
def take_attendance(list_of_students):
"""Use at beginning of class to record which students are present"""

  for student in list_of_students:
    # Do the following for each student in the list
    answer = call_student_name(student)

    if answer:
      mark_present(student)

    # Do nothing if no answer

  # Now we're done calling the initial list
```

[6] For a discussion of this see Ref: *ALGORITHMIC HARMS BEYOND FACEBOOK AND GOOGLE: EMERGENT CHALLENGES OF COMPUTATIONAL AGENCY*. (2015). *ALGORITHMIC HARMS BEYOND FACEBOOK AND GOOGLE: EMERGENT CHALLENGES OF COMPUTATIONAL AGENCY* (pp. 1–16).

```
missed_students = ask_if_anyone_was_missed()

if len(missed_students) > 0:
   for student in missed_students:
      mark_present(student)
```

Note that `mark_present` and `call_student_name` are two other functions which do what they're named.

5 Databases

Data of the sort which concerns us is stored in databases. There are many different formats, from single spreadsheets to complex relational databases with hundreds of tables.

In a relational database, the information will be broken up across a series of tables with another set of tables connecting them together.

[ToDo: Flesh this out or find video/explanation to suggest]

This is important because abstracting out the parts of the data (we separate the thing, the property, etc) allows us to connect things in novel ways. For example, we can query the database to learn new stuff. This enables us to more easily do data-mining.

6 Data mining

[ToDo: brief overviews of how some data mining techniques work]

7 Main Questions

Companies have collected data on their customers forever. Many of the hot data-mining algorithms have roots in statistical techniques that have been around awhile. Why are we so worried about this stuff now?

There are probably a lot of factors. One major set of changes involves the drastically declining costs of storage and computation. It used to be that if you wanted to keep data on something, the benefit of that data needed to outweigh the costs. Now, the marginal cost of storing data and processing it is basically trivial. In many cases, there is very little financial reason not to capture all the data you can and store it for an unlimited amount of time. [7]

Another factor is the availability of extremely individualized data. The internet enabled this to some degree. But the rise of social media and the ubiquity of smart phones makes it possible to gather a detailed profile of every potential customer.

Indeed, the wealth of very granular data has given rise to companies who specialize in what used to be called Knowledge Discovery in Databases (KDD). These companies aim at

> discovering non-trivial new insights in existing datasets, insights that cannot simply be observed in datasets or follow automatically from datasets, but insights that have to be extracted or generated since they do not 'lie at the surface' [Sax 27]

Nowadays, we call this big data.

To paraphrase an expert on the B.I.G., big data, big problems.

7.1 (Q1) When may an ethical company profit from the use of personal data?

As Sax notes

> big data's entrepreneurial potential resides in the fact that advanced mining techniques can extract/generate unanticipated, non-trivial, new, and (commercially) interesting insights. [Sax 27]

If that's true, then as he writes

> Big data's entrepreneurial potential is equally dependent on the legitimacy of the appropriation of these newly extracted/generated insights by commercial parties [Sax 27]

This is roughly our first main question:

> (Q1) When may an ethical company profit from the use of personal data?

This extends to insights derived through machine learning and other analytical techniques done on users data.

Q1 Answer 1: When they own it

[7] Check out how cheap Amazon Web Services is: https://aws.amazon.com/pricing/

Q1 Answer 2: When we let them

7.2 (Q2) Harms of informational privacy violation

> (Q2) When is a misuse of personal data significant enough to warrant moral condemnation, regulation, criminalization, or other forms of coercion to prevent?

Q2 Harms of informational privacy violation

7.3 (Q3) Responsibility

> (Q3) How should we assign responsibility when people are harmed by algorithmic uses of personal data?

Q3: Responsibility