(OHMYGAWD this is a lot :p)

[Terms and Conditions Key Words and Concepts in terms of data ethics](https://docs.google.com/spreadsheets/d/1qr5kZYCPlnYd1PlwAiu6xgzB22JUzYpvOTiEMgUGfIo/edit#gid=0)

MAIN FUNCTIONS

* Site that generates 7 principles: what does a company do with its data, and is it good?
  + Checkmarks for Privacy, …, Data shared with 3rd parties
  + Full transparency of what’s embedded in Terms & Conditions. -> translate the legal language.
* START: scraping the website: scrape *Terms and Conditions*
  + Text processing guide:
    - <https://www.kaggle.com/amr009/ethical-data-collection-for-financial-news/notebook>
* Pick 2 companies’ *Terms and Conditions* -> take notes within each of these **principles** -> tomorrow, regroup and share key words and how we want to rank them.
  + <https://www.facebook.com/terms.php> <- Prashi
  + <https://twitter.com/en/tos>
    - <https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/asset-download-files/TheTwitterUserAgreement_1.pdf> <- Prashi
  + <https://policies.google.com/terms?hl=en-US> <- Marco
  + Microsoft, Amazon <- Dean
* Later on: Chrome extension pulls past results from site
* (Visualize site data across sites <- extra feature)
* Before a user visits a new site, use NLP to read privacy agreements, and provide a mechanism to continue to the site.
* Improve the knowledge base of users, so that they are notified of privacy based on the following **principles**:
  + Privacy <- what do we really mean by ‘privacy’, what are our criteria for *good* and *bad*?
    - Train our model based on criteria
  + Consent
  + Consequences
  + Data Collected
  + Primary usage of data collected
  + Secondary usage of data collected
  + Improper access after collection
  + Errors in data handling

Ethical Data

Ethical Data- It's about studying and evaluating moral problems related to data.

moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).

Data Ethics focus on third-party practices with individuals’ data, while information ethics span more broadly to media, journalism, and library and information science.

Principles-

Fairness, Reliability, Safety, Transparency,Privacy and security, Inclusiveness

forget about google analytics and what according to you are the current practices that you consider as unethical?

Human rights can do this debate.

Tech literacy is a must.

Privacy - use of user data, prediction of sensitive attributes (e.g., an employee churn prediction model will often flag all women age 28-35, the most common age to get pregnant)

Consent - is it ethical to use my medical data to improve models to read CT scans? What level of information do I need to reasonably consent to that?

Security - adversarial attacks, data security

Layoffs - is it ethical to outright automate people's work? An entire industry?

Bias - literally google "AI Bias"

Transparency - google explainable ai. Interpretability, auditability, data source traceability, how it's updated, etc.

Medical research are also having user's data.

Companies having user's data by purchasing location data.

People are not happy with sharing data but does it make it unethical?

Like, the difference between “my doctor has a bunch of data on me and can prevent me from getting sick” and “amazon has a bunch of data on me and can warn me when I’m about to run out of diapers” is a matter of degree, but ultimately the value exchange is there. if you can improve customer experience quantifiably, isn’t ANY data collected in the course of doing so ultimately “ethical”??

Any ideas with this dataset <https://www.kaggle.com/amr009/ethical-data-collection-for-financial-news>

How human rights and data ethics collide? consent is needed?

what is considered to be public data for research work?

IRBs

Twitter tweets used for research? 70% dont have any idea

users thought it's not allowed.

twitter has in privacy policy that researchers can use their tweets.

* Is their data shared with advertisers?
* These are the key points of how your privacy is being handled when you want to sign up for a page or site.
  + Either direct copy and paste into our tool, or our tool could be a Chrome extension if we run analysis on the 50 most popular sites.
  + Education
* NLP for scraping privacy policy
* Legal documentation -> legal docs are standardized -> look at conventional format

Considering Blockchain and decentralized applications

Let's understand Blockchain

Blockchain uses peer to peer technology to send payments or data from one entity to other without any involvement of financial institution. Don't need to trust any 3rd party intermediaries to complete a monetary transaction. It does it like this to make sure that data is completely secured and it is efficient without excessive cost and time. Transactions are borderless (sending to someone in India, Japan or Germany is no different from sending to your neighbour)

Blockchain aim is to create a digital form of cash and trusted form of currency.

Blockchain is a way to store data securely.

Basically, what happens is your data is divided in blocks. These blocks are connected in a chain, with each block having a security code that depends on the previous block.

This makes hard to manipulate data. If there are 10 blocks in the chain and you manipulate block 1, then the security code for block 2 is wrong. So you need to manipulate block 2, but then block 3 is wrong. So you need to keep manipulating all the way to the end.

blockchain is a glorified spreadsheet (think: Excel with one table). It's a new way to store data.

In traditional databases there’s usually one person who’s in charge, who decides who can access and input data, who can edit and remove it. That’s different in a blockchain. Nobody’s in charge, and you can’t change or delete anything, only view and input data.

Since blockchains can have a tremendous amounts of blocks (527125 for bitcoin right now), the fact that you have to go through everything makes messing with stuff a hassle.

Two types of Blockchain-

1. Permissioned Blockchain- It's a centralised authority which allocates responsibility to individuals to operations of blockchain. limited people can join in.

2. Permissionless Blockchain- Anyone can join network and take part. Anyone can operate as full nodes and start mining. Just like Ethereum and Bitcoin.

Now what's bitcoin?

Simply, a digital currency that uses blockchain.

if anyone wants to study more <https://github.com/bitcoinbook/bitcoinbook>

what's relevant with decentralized applications?

Trustless business models which are not possible with our traditional user-to-server architectures. For instance enforceable contracts, tokenisation, transparent voting, peer decision making algorithms and more.

Ethereum major role in gaming

Some existing games that use Ethereum Blockchain- KryptoWar, Gods Unchained, EtherKingdom, CrytoKitties

<https://cryptoguide.ch/games/eth/ethgame3.html>

what is meant by 100% decentralized game?

It means all data used to run the game is stored on Blockchain(basically we don't have any database)

In KryptoWar- Grab the army names, soldiers count, battles results directly from the Blockchain using Web3.js (the Ethereum Javascript API).

With centralized applications:

If you want to build a website or an app right now and want to store some data like usernames or recipes you will need to have your own database that will let you store and retrieve data. This is very effective but it has some drawbacks:

You can be hack and get all of your data stolen

You can lose your data because of simple mishandling, a bug, an hack

It is centralized which means that you can be the only person/company accessing those data (makes collaboration difficult)

With decentralized applications(DApps)

If you build a similar website or app on the Ethereum Blockchain:

You won't be hacked, it's impossible for a hacker to delete data from the Blockchain (if you don't allow it in your smart contract)

Anyone could build services on top of your smart contract (which could open up a new era of innovation and interactions between companies)

You can create a true ownership of data for your users

Game on the Blockchain, any decentralized app could connect to our smart contract with Web3.js and create another game for the owners of our armies.

DESIGN OF SMART CITY SYSTEMS FROM A PRIVACY PERSPECTIVE

Because supposedly over 50% of our global population now lives in cities, and because cities are complex systems characterized by a massive number of interconnected citizens, businesses, communication networks, and services and utilities, Smart Cities need to combine and balance factors including: economy, mobility, environment, people, living and governance, built on the smart combination of endowment and activities of self-decisive, independent and aware citizens.

We need to prevent smart cities from invading citizens' privacy, because if we only knew what some of our citizens were up to then we would literally not be able to function as a society.

The aim of the research is to gain an understanding of citizens' perspective on privacy issues, treating citizens as rational beings and yet making sure that power is not allocated to the unearned. I mean, we really want to make sure that information collected is used for its original purpose. Six key dimensions have been identified: (1) collection, (2) secondary usage, (3) errors, (4) improper access, (5) control, (6) awareness. Essentially, large amounts of Personally Identifiable Information are collected, used for secondary purposes, and we've got to pay attention to the degree of effectiveness and the margin of error between what a person thought was true and what actually happened with regard to protections against improper access; their level of awareness of the truth versus the degree to which they perceive their adequate control over personal information.

This guy called Langheinrich apparently contributed to the area of privacy by design by developing the following practices (you know he did):

\*This is what we’re doing

**\*Notice (notifying the users when you collect their data)!,** choice and consent (explicit consent, come on people), proximity and locality (they've gotta be there when you're collecting it!), anonymity and pseudonymity (we should provide our own service in the event that they actually don't consent to the collection of data from their device.).

**\*need a security mechanism before they go on site**

Oh! And there are more. Security: there should always be security mechanisms providing adequate protection for collected data. Access and resource: Access to the user's data, once collected, should only be allowed to authorized persons.

So this research project is about implementing the Action Design Research methodology.

It revolves around a 1000-citizen survey regarding citizens' perspectives on large scale "real-life" experimentations of intelligent acoustic solutions providing "situational awareness" by using audio monitoring in combination with Internet of Things (IoT) technologies, via a deployment of Audio Processing Units (APUs) in the targeted in-door and out-door environments as complementary to intelligent sensors - the APU consisted of a microphone and an embedded processing platform that continuously "listened" to its environment and analyzed the sound locally.

They collected surveys - familiarity with the concept of audio monitoring, concerns about the collection of personal data, secondary use of data and improper access (when the organization collecting the data shares the data with other organizations), control and awareness (concerns about control and awareness of their audio being collected), usage scenarios...

So the design principles for privacy protective smart city systems are, based on what citizens were not overly positive towards in terms of audio monitoring:

General placement of sensors, no personal data storage, stream data, sensor systems in solitude.

These are some of the factors we might look for in our projects. **Help improve their knowledge base of these factors.**

SERIOUS GAMES IS ACTUALLY A CONCEPT

Serious games are games whose primary objective is NOT fun or entertainment, rather learning, practicing a skill or a primary purpose.

Eg a therapy game for a children's rehab centre

Our primary purpose/objective is privacy and security[ethical data]

<https://www.hiig.de/en/serious-games-as-a-tool-for-privacy-by-design/>

The first concept, "Who's Sherlock" is a Q and A game that lets its players experience how easy it is to find information about oneself and others on the internet, and how sensitive this kind of information can be in different contexts.

The second game concept, "Privacy Rush", illustrates the moment when an individual wants to connect with a public wifi system and seeks to inform them prior to when they actually use the insecure connection, about the privacy risks.

The third game concept is a card game: "Data Trade" addresses the complexity of today's data economy by demonstrating, from the perspective of a day trader, how different data sets can be combined, leading to a higher price on data markets because they generate more information about individuals. Do you know what they are talking about? They are talking about inference: just as an example, multiple pieces of anonymous data might be re-identified to re-identify individuals. *Based on the combination of insignificant data, we can achieve significance.*

**\*\*\*\*\*\*\*\*We could visualize their site data across different websites to show the sum total**

The fourth game concept, "Nothing to hide", lets the game players experience that the "nothing to hide" argument is wrong. If anyone can find more information about this game concept then let us know, because I want to know specifically what they did.

The fifth game concept, "Smart City", is a simulation game targeting political decision makers who are responsible for urban development decisions. At the end of his or her legislative term, the player (who takes the perspective of a city mayor making political decisions) will be measured pursuant to the future scenarios simulated in the game.

The sixth game concept, "Pieces of Data", seeks to teach its players how to *manage third-party access rights*, i.e. by app providers, to their smartphones. "Taking the perspective of an activist in a state that is becoming more and more repressive, the players learn how data collected by their smartphones can harm them, and even other parties - and how to protect themselves against such threats.". Apparently this game won! We could make a similar game, which would be hilarious!

TERMS, CONDITIONS, COOKIE POLICY REVIEW

FACEBOOK

1. We track editorial changes to analyses and updates to a point's status and display the previous versions here as part of an effort to promote transparency regarding our curation process.

Tracking pixels used in service-to-user communication

2. We use cookies if you have a Facebook account, use the Facebook Products, including our website and apps, or visit other websites and apps that use the Facebook Products (including the Like button or other Facebook Technologies).

Facebook uses, pixels and local storage in order to gather information about you, your device, your browser cache

Facebook also uses cookies for advertising purposes.

3. Information collected by these third-party services is subject to their own terms and policies, not this one. It's in data policy

Facebook informs users that its privacy policy does not apply to third party websites

4. Product research and development: We use the information we have to develop, test and improve our Products, including by conducting surveys and research, and testing and troubleshooting new products and features. Facebook Data Policy

Facebook uses your data for many purposes

5. You give us permission to use your name and profile picture and information about actions you have taken on Facebook next to or in connection with ads, offers, and other sponsored content that we display across our Products, without any compensation to you.

Users identity is used in ads that are shown to other users

6. Facebook uses personal data for advertising

7. Users are not allowed to use pseudonyms, as trust and transparency between users regarding their identities is relevant to the service.

8. Personal data given to third parties.

Sharing with Third-Party Partners We work with third-party partners who help us provide and improve our Products or who use Facebook Business Tools to grow their businesses, which makes it possible to operate our companies and provide free services to people around the world.

We share information about you with companies that aggregate it to provide analytics and measurement reports to our partners. Partners offering goods and services in our Products. When you subscribe to receive premium content, or buy something from a seller in our Products, the content creator or seller can receive your public information and other information you share with them, as well as the information needed to complete the transaction, including shipping and contact details.

9. Facebook stores your data whether you have an account or not

Facebook uses cookies and receives information when you visit those sites and apps, including device information and information about your activity, without any further action from you. This occurs whether or not you have a Facebook account or are logged in.

10. gathers information about you through third parties

Data Policy- We also receive information about your online and offline actions and purchases from third-party data providers who have the rights to provide us with your information.

TWITTER

1. uses personal data for advertising.

If you like, Retweet, reply, or otherwise publicly engage with an ad on our services, that advertiser might thereby learn information about you associated with the ad with which you engaged such as characteristics of the audience the ad was intended to reach.

2. Gathers information of users through third parties

3. You can download certain account information, including your Tweets, by following the instructions here. On Periscope, you can request correction, deletion, or modification of your personal data, and download your account information, by following the instructions here.

You can retrieve an archive of your data

4. Pseudonyms allowed

You can also use Twitter under a pseudonym if you prefer not to use your name.

5. option to opt out of promotional communications

You can control most communications from the Services

6. Freedom to share content- You can choose with whom you want to share your tweets.

one more option- you can choose who can comment on your tweets

7. Users should be of a certain age. you must be at least 13 years old

8. prohibits users from attempting to gain unauthorized access to other computer systems

9. Revisit terms and conditions from time to time. Notification sent to users

10. However, some of our services may not function properly if you disable cookies.

11. collects your IP address for location use

12. may also collect other personal data from you, such as cookie identifiers or your IP address.

13. may stop providing service

may stop (permanently or temporarily) providing the Services or any features within the Services to you or to users generally.

14. tracks you on other websites

15. have the right to leave this service at any time

DUCKDUCK GO

1. IP addresses of websites visitor not tracked

2. We don't need to register

3. DUCKDUCKGO doesn't track users.

4. Cookies don't contain info that would identify users

* We need to convert URL pages data to text data (maybe json) by scraping web pages
* Data exploration and data cleaning
* Clean and Prepare data with NLP Techniques
* Transforming our text to variables. Training our model.
* Binary classification(if we are going for yes/no)[confusion matrix can help here]
* Testing our model by various diff algorithms [Logistic regression for binary classification]