**Matthew Gilmore - Revised Writing 1**

How many people actually read the terms of service and privacy policy of websites like Facebook or Twitter? On both websites the links to the pages describing these policies are in plain text beneath the forms used to create an account. The user agrees to all the terms just by signing up. And even if a user reads the terms of these websites, each policy declaration is many pages of legal terms designed to cover the liabilities of the social media company. Most people are not lawyers and do not take the time to read the legalese before creating an account. Currently, there is no system for users of websites to negotiate just how much data they want to give away, nor is there an effective way to keep companies accountable for their actions.

To address these problems we propose a system of data guardians to serve as advocates for consumers with the use of blockchain technology to keep privacy agreements public and verifiable. We will use the phrase ‘data guardians’ to refer to public, private, and non-profit organizations that help fill the gap of technical knowledge between companies and users. Before creating an account with a new social media site, users will be asked to create a policy with a data guardian of choice. Data guardians will do the legal and technical work when it comes to learning the fine details of privacy agreements. All users will have to do is read the data guardians’ simplified descriptions of what the website in question is asking for. These descriptions will be tailored for user understanding and will be stripped of the legalese that only benefits companies.

However, this system still requires incentives. First off, this system does not require funding from private investors. In order to get both companies and data guardians to participate, we propose a system of credentialing similar to how financial institutions are regulated today. Therefore the government likely needs to set up a nonprofit organization that is shielded form monetary concerns. Also, data guardians would keep a record of their interactions with companies and how the companies handle user data. There will be a system to trace how data is used each time and the ability to reference the policy allowing said data usage. Data guardians will be encouraged to use this information to give public scores to each company in how they handle user data. Much of this information will be public. By creating an official network of data guardians, we will help users more properly discern which companies are worth interacting with.

Next, one of the technical challenges of this project involves the implementation of a public blockchain technology. Many may know this term from its applications in cryptocurrencies, but it also works as a way to keep all parties in our system accountable. Blockchains serve as a public, unchangeable ledger holding a record of all transactions. Each block is cryptographically linked to the previous in such a way that any modification of the block would break the chain and be a noticeable violation. Furthermore, the user would sign every transaction in order to verify it. There are also several types of consensus algorithms that allow all people to maintain the same copy of the ledger. All together, finding the right technical functions to keep the ledger secure will create a place to hold all contracts between users and companies.

In total, our data guardians will shift the balance of control over personal data into the hands of users and not private companies. Through new technologies and laws, we will create incentives for good behavior among all stakeholders and encourage a more just use of the internet.

**Daniel Marley - Revised Writing 1**

The Need for “Data Guardian” In 2019 digital ad revenue topped $100 billion for the first time1, the reason being their efficacy. Targeted ads become more effective each year as we continue to create and give away data, revealing more about who we are individually. As of last year, 2.5 quintillion bytes of data are being generated every day2. This data can range from basic public information we input ourselves – our name, our age, location – to usage data we may not realize we are signing away – such as shopping habits, frequently visited sites, travel habits, job fields, etc. Companies know the value of this, as American companies alone spent $19 billion this year acquiring it3. This is the cause for our project, “Data Guardian”: a means for a user to track the data the user shares with companies.

Now, we achieve this goal as “Data Guardian” as a service upkeeps a ledger of company reported data usage. In practice, before an individual creates an online presence, he will create an account with our service. We will then allow him to create a custom data policy, mandating what data companies have access to, specifying a lease for how long the company may use said data, and requiring companies to report when they use that data. Then, when the individual creates an account on another service, such as Facebook, the user links their new Facebook account to their “Data Guardian” account. Facebook would agree to the Data Guardian policy terms in exchange for access to the users data; as part of that agreement, Facebook then reports to the Data Guardian any instance in which it uses that user’s data. The Data Guardian then reviews this report and adds it to our ledger, for the individual to view.

Shifting from this idealized view, there are several technical challenges within this project, such as the design ledger and the means of maintaining its security and preventing false reporting. Another technical challenge is how to build such a system with the ability to scale, as the amount of internet users is only set to grow; Blockchain is built for growth however, as we will not store any of the data beyond the basic policy tables per user and reported data-usage from companus. Lastly, there is the challenge of how to account for company participation; companies, however, are looking to re-establish trust after several public exposures of unknown data use – such as Cambridge Analytica in 20164. This would be a means for them to operate, still with a profitable data centered model, but without heavy regulations that are sure to come if the industry may not regulate itself. Lastly, the project would be funded from the government or an NGO, dedicated to watchdog organizations.

1. “How Much Data Do We Create Every Day?”, https://www.forbes.com/sites/bernardmarr/2018

/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should- read/#70f458160ba9

2. “Digital ad revenue in the US surpassed $100 billion for the first time in 2018”,

https://www.cnbc.com/2019/05/07/digital-ad-revenue-in-the-us-topped-100-billion-for-the-first- time.html

3. **“**The WIRED Guide to Your Personal Data (and Who Is Using It)”, https://www.wired.com/story/wired-guide-personal-data-collection/

4. “Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens”,

https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

**Jonathan Minkin - Revised Writing 1**

When we interact with companies on the Internet, we often give these companies access to our personal data. The company will ask for some basic account information in its sign up form, and will display some disclaimer prior to confirming a user’s registration. This disclaimer contains a long legal document which lays out the terms of use of that company’s service, including how it can use your personal data. For most, this document is too complex to understand in a reasonable amount of time. However, for most services it is necessary to agree to the terms of the document to use the service at all, so many people are likely to accept the terms of the document without reading it. This creates a problematic system of data usage. Since the company is writing the terms and conditions, it gets to decide how it uses your data. It may also not disclose all the ways in which it uses your data. Furthermore, there is no way for a user to verify if a company is properly following their terms when using your data.

This problematic data usage pattern requires a solution that returns control to users. Thus, we propose a system that changes how data usage policies are created and enforced. Our system introduces an entity which we call a “Data Guardian.”A Data Guardian is a non-profit entity that provides a website where a user can go to create a data usage policy before they sign up for a website. The data usage policy will use common language to describe what data the company has access to, how long its access is valid, and how the company is allowed to use the data. A data usage policy can be provided to company when a user signs up for the company’s website. All companies that use personal data will be legally obligated to report all data usages to a Data Guardian, along with the policy that was provided by the user. This data usage will be logged using a publicly available ledger called a blockchain. This blockchain will be unchangeable and cryptographically verifiable, so that no one can tamper with data usage history and anyone can verify that a company is properly using data. Data Guardians are meant to be funded by the government, and should not exist to make a profit. The major technical challenges in this system involve the maintenance of the blockchain and the creation of data usage policies. The blockchain will be stored on a large peer-to-peer network of computers, and all computers need to agree on the state of the blockchain at any time. This will require an algorithm that coordinates all updates to the blockchain. Furthermore, data usage policies will have to be cryptographically verifiable, so that no one can forge a user’s policy and then use it on the blockchain.

Our system aims to be easy to understand, and should not require extensive legal or technical knowledge to use. Consider a situation where a user wants to sign up for a popular online shopping website. The user wants to use the site because it will give him helpful shopping suggestions based on past purchases on the site. However, the user does not feel comfortable allowing his information to be used beyond this website. This user goes to a Data Guardian, and creates an account with the data guardian. The user then creates a policy for this shopping website which includes a clause stating that the company is not allowed to release shopping data to any third party. The user then goes to the shopping website, and provides this policy during account registration. After several months of shopping on the site, the user sees an advertisement on a different website for a product he recently bought. The user clicks on a box on the advertisement, which shows that an advertising company got his data from the shopping website. The user can report this to the Data Guardian, which will serve a takedown notice to the shopping website for violating the data usage policy. By having the Data Guardian system in place, the user was able to hold the shopping company accountable for their data misuse.

**Comparison with Similar Products**

Our product largely has no relevant competition. Data privacy is a huge problem throughout the world, yet the technology and global markets have made it very difficult to come up with a specific system to approach it. Companies exist outside of countries and can often move to avoid laws in one country in favor of more lenient ones in another. These companies are also typically giants of the industry with many resources at their disposal to fight any kind of regulation or competition that attempts to do business regularly. Therefore, any system that tries to go against these forces will have to be a serious undertaking of a large government structure that has the resources to enforce its laws.

With this in mind, the only major project similar to our product is the European Union’s General Data Protection Regulation[[1]](#footnote-0). The law strengthens consent laws for users in order to protect their data usage. It seeks to restrict the legalese of privacy agreements in order to favor user readability. Users will also have the right to access their data or to tell the company to remove all data on the user. This data will be sent to the user upon request as well as any information on how it was being processed. The law also requires the design of strict data privacy structures. Companies are required to create their systems with certain securities built in to prevent leaks of personal data from the start. Lastly, certain companies that deal in large amounts of sensitive personal data would be required to hire a Data Protection Officer. This DPO would be a member of the organization who keeps the appropriate records to meet the requirements of the law. Many of these stipulations will include heavy fines if they are not followed.

This law has some ability to work for several reasons. First, it addresses the problems stated above in its scope. The European Union is a large organization of many countries that would have the resources to go to court with a big technology company if necessary. A smaller country might not be able to do this. Furthermore, the law applies to any company that uses the data of EU citizens, not just companies based in the EU. For these reasons, the law has some possibility of success. It also has started the process of creating dedicated professionals to hold companies accountable for data privacy in the form of Data Protection Officers.

However, our product would still be able to succeed on its own. First, no comparable system currently exists in the United States. Therefore there would not be any competition with our product. Ours is also different in important ways. First, ours does not rely on any centralized database to store the necessary information. Powerful companies will find it difficult to attack our decentralized storage. Also, our system has data guardians. The DPOs of the EU’s system are rare and susceptible to monetary concerns as they work for the companies in question. Data guardians will be independent and all of their work will be publicly verifiable in the ledger. Therefore, the existence of other similar products will not limit ours.

**Target Audience**

The Data Guardian system is targeted at those who voluntarily give their personal information to an online company. Currently, a person signing up for a website will likely agree to a long list of terms and conditions laid out by the company that runs the website. For most users, the legalese or technical language of the terms and conditions is too difficult to understand. Users in this category are likely to accept the terms and conditions without consulting anyone who can understand the policy. On the other hand, those who can understand the policies may not be comfortable with the terms under which their data can be used. However, agreeing to the terms and conditions is usually a requirement to use a service, so users in this category are forced to consider if their desire to use a service outweighs the discomfort caused by the data usage policy. Our system aims to reduce these two groups of users into a single group that is well-informed about how its data is used. Users will not need technical or legal knowledge to use the system, and anyone can customize data usage policies for their level of comfort.

Our goal is to make this system available to as many users as possible, which is why the system will be free to use. This system is meant to be funded by the government, and thus does not exist to make a profit. Furthermore, since this system is meant to be accessible to the general public, it is important that anyone can understand how to use it. One of the key issues that our systems aims to fix is that it is difficult for the average person to understand the legalese in the terms and conditions of many sites. As a result, our system will present users with data policies that are written in common language. This will allow users with limited legal and technical backgrounds to understand how their data will be used. We will also provide pre-made policies for popular websites, so users do not have to spend time creating a custom policy for every site. Users who opt to use a template instead of a custom policy can still trust that the policy they use will serve their best interests, as it is provided by an organization that does not stand to profit from their data. The Data Guardian system will be easy for any person to understand, and will not require a significant time investment to learn and use.

Our goal of accessibility is important because the success of the Data Guardian system will be dependent on participation from users and companies. Participation from companies will come from legislation that is meant to be introduced alongside the Data Guardian system. Companies will be required to offer an easy way of integrating with the Data Guardian system, and will be required to notify their users of the availability of the new system. Participation from users will come from government-funded public awareness campaign to inform users about the existence of the system. The main source of information will be a government website which describes, in simple terms, what the Data Guardian system does, and how users can start using it. Creating advertisements in all major mediums, especially on the Internet, will help create awareness for this system. Furthermore, users that already have an account with one of the affected sites will receive a notification from the site alerting them that the Data Guardian system exists. This will prompt users to visit the Data Guardian website out of curiosity. This will prompt users to visit the Data Guardian website out of curiosity. Similarly, new users to any website will be prompted for a Data Guardian policy id when they sign up, and will be able to open the Data Guardian website in a new window to learn more. Through legislation and advertising, we will promote both corporate and user participation in the Data Guardian system.

**Societal and Global Impact**

Now, shifting to the broader societal and global impacts our project could have, the changes cannot be understated. For the first time within the digital era, consumers would know with absolute certainty where their data is and how it is being used. Our project will also allow users to identify untrustworthy companies, who repeatedly break data agreements, before they share their data. Companies who properly report and properly use data will be able to build trust with the consumer. Important and often useful tools, such as targeted ads, will seem less invasive to consumers if they are able to see exactly how their data was used to generate said ad. Perhaps most importantly though, our project would implement the legal standard of data being owned by an individual and leasable to companies for a period of time in exchange for a service. This is a huge shift from the current state of data ownership, once a company obtains your data it may store it for as long as it pleases and use or sell it how it wishes.

Off of that point, with widespread use of our service amongst consumers, the average individual would certainly become more conscientious of what data they give away and to whom. If it were to be used widely amongst companies, it would pressure non-participating companies to join as well so that they may reach a competitive bar of data safety other services guarantee. All of this, of course, only functions with a fair amount of government regulations to support our project.

The first and foremost of these regulations is the legal establishment of data as owned by a private individual and merely leasable to companies; some penalty, most likely a financial one, will also need to be established for breaking one of these lease terms. Additionally, there would need to be a legal impetus for companies to participate in our service: this may be a financial incentive such as a tax credit to create some momentum amongst joining companies or it may simply be a legal mandate to participate. Lastly, there would need to be added regulations on who would be able to work within Data Guardian, maintaining the ledger of data accesses, as it should be insulated from corporate influence.

On another note, it is worth asking if such a novel implementation of technology could be misused or have potential for exploitation by bad actors, but the simple answer is no. The goal of our ledger is transparency and for the public to be able to view when companies are using specific data parameters. We do not make any of the data itself available as we do not have access to it. Additionally, internal workers of Data Guardian may only modify the ledger, so there is no danger of data being falsified.

Lastly, it is notable to consider the global impact of this service saturates the US. It is possible that companies would wish for uniformity and would work to deploy this system of data assurance to their customers internationally. Though, it is equally possible that companies will prefer to stick to the more profitable status quo of data handling internationally. Only implementation of our project would lead to a sure answer.

**Technical Innovation**

The use of blockchain technology is one important way that our project does something new by expanding upon previous technologies. A blockchain is a system for storing data that needs to meet several requirements. The data in this system must be publically viewable and many different computers must maintain it without the use of one central authority. Thus, there needs to be a way for users to agree upon what data is in the blockchain and ensure previous data is not tampered with.

Most people are familiar with blockchain primarily through its use in cryptocurrencies, most famously Bitcoin. Bitcoin solves the problem of protecting previous data by grouping monetary transactions into a block and then linking each block to the next in a list. Each block has an identification number that users adjust so that when the entire block is run through a hash function, it results in a specific type of key. This hash function and key can be checked to ensure no tampering was made to the items in the block. As long as most of the users keeping track of these blocks are not malicious, the system will keep and add new data in a way to prevent tampering.

However, our system must work in a slightly different way because we are not implementing a monetary system. When using a blockchain, all computers that participate in the network must agree on the state of the blockchain. The method that is used to achieve consensus varies based on the blockchain. For example, Bitcoin uses a system called proof-of-work, in which all computers are continuously running a hash function using the latest block to try to achieve a particular value. The first computer to find the value “wins” the block, and is rewarded with an amount of Bitcoin. However, proof-of-work is energy inefficient, and participants in the network expect a reward for the computing power and electricity they use to participate. Since our blockchain is not used for currency, it is difficult to offer a reward. Furthermore, we would like to minimize the amount of energy that is consumed using our blockchain. Therefore, we plan on using a different consensus algorithm called Raft. In this algorithm, all computers participating in the network will elect a “leader” machine, which is then responsible for handling all write requests to the shared blockchain and distributing these writes to all other computers in the network. The algorithm is fault-tolerant, which means that it is able to properly sync all computers in the network in the event that any computer fails. Furthermore, our system will also be closed, meaning not just anyone can be a node in the system and add transactions. Data guardians will have to go through an approval process through a non-partisan government agency before they are allowed to enter the system. This solution allows us to use the aforementioned consensus algorithm without the possibility of bad actors simply making up transactions once they are lead nodes.

**Technical Feasibility**

In order to build our blockchain, we will need to use an implementation of the Raft consensus algorithm. This will allow all data guardians to coordinate the state of the blockchain at any point in time. The implementation we plan to use is called etcd, which is a distributed key-value store that coordinates network participants using Raft. Furthermore, there is a feature in etcd in which any participant is able to “watch” a key for any updates. All data guardians in the system will participate in the same etcd network, and will all watch the same predetermined key for updates. Any writes to the blockchain will be represented as a string, and will be mapped to the predetermined key. Every computer in the network will receive an update that the key has been updated, and will be able to see the transaction as a string. Each computer will then update its local copy of the blockchain with the transaction. This system will allow for consensus among the data guardians that will run the blockchain.

One of the important features of our system is the security it provides to users. Each individual contract between a user and company must contain the signatures of both parties involved. However, the personal information of the parties should not be exposed. By signing a transaction, the identities of each party should not be traceable back to the actual users. In order to implement this kind of security, we need a type of digital signature that can be signed in a way that proves the correct party is agreeing to the contract by being publicly verifiable, but not revealing the real identity of the person. This can be done with a public key, private key pairing. A user needs to be able to securely generate a private key that is not published and a public key that is. This private key will be used to sign transactions by creating a signature dependent upon the message. The signature can then be checked using the public key but cannot be forged without the private key. Thus, users will need to publish a public key as their identity, but they are not required to link their real identity to this public key. This security can be implemented using the Java Cryptography library. This library has an API that allows programmers to implement the key generation and signature functions without having to create it on their own. Java even has a built in provider that implements the specific algorithms that the functions run on. Therefore, providing security for users is possible using existing tools.

Customers of a specific data guardian need not know all of the technical specifications of the system. Ideally, when interacting with a data guardian, the customer agreeing to a contract with a company will have the specifics hidden from view. This will allow any customer to set up an account with ease without needing technical background first. In order to set up such a system, we will use creative website design to ensure customers of data guardians have an easy time setting up their accounts and creating contracts. To create the websites, we will use existing tools such as Angular. Angular is a JavaScript framework that allows for the creation of dynamic websites that change to match the user’s needs. In order to store the information of users we will also use PostgreSQL for our database. This allows for the creation of relational databases to store and then retrieve user information, such as usernames and passwords. Both of these tools already exist and can be quickly integrated into our overall project.

**Cost, Risks, and Risk Mitigation**

Shifting to the topic of costs, it is important to consider the storage we will need for user information and our ledger, the servers we will need to run our web application, and the length of the code itself. First, considering our storage, we will only be storing usernames and passwords for each user and then the length of the blockchain storing our ledger of company data reporting. Within the development phase, we will be using open source free software to locally manage user data tables and will hold a local copy of the blockchain. Looking forward, from early development to basic deployment it is likely that we will utilize a scalable database tool such as AWS’ database, which charges only $1.15 per 10 GB of storage.

Focusing now on our web application, we will be using Angular as our development framework with supporting free libraries, both for the front-end design and back-end creation of services. None of these resources require any costs. The running of these services will also start locally and will be tested locally,but will likely shift to a service such as AWS lambda for service hosting and AWS S3 services to host our web application. AWS lambda allows for up to 1 million service requests per month before charging and AWS S3 reports that most websites, if they go over free limits at all, only cost $1-$3. This will be highly manageable for our initial development purposes.

Turning to the length of what we will actually develop, it should again be noted that we are utilizing an Angular framework with supporting libraries that will assist the UI, cover cryptography, and part of our blockchain management. The additional services that we will be developing are as follows: Company Reporting REST API, Company Registration, Blockchain Lookup, Login/Registration Component, Legal Flag System, Lease Management, and Policy Creation. From past professional experience, I’d estimate that about each of these services will be around 400-500 lines of code, being an upper bound of 3500 lines for the web services.

Pivoting to the user interface, it is again noteworthy that we are on an Angular framework with supporting UI libraries from Google. Based on previous extensive experience with these tools, it is likely there will not be more than 500 lines of CSS code for the interface. In terms of HTML, each page will likely average from 200-300 lines with around the same amount of supporting Javascript. With 6 main user interface pages being considered – Login and Registration, Policy Management, Auditing the Blockchain, Report Data Usage, Legal Review, and Lease Management – this places the total estimate for the web interface at approximately 4100 lines.

Lastly, in discussion of our milestones for our project as we develop, we have determined five benchmarks and their criteria. First, we have our Bootcamp Phase, where we will achieve Basic Running Backend Service and Front End, a functioning database, and will define and implement our blockchain structure and storage. This will be followed by Phase 1: a viewable blockchain within our web app, completed user registration, login, and account recovery, and a completed company reporting service, policy creation service, and blockchain lookup service. Next will be Phase 2, where we aim to complete the legal user view, the legal reporting service, the private user view, the consensus algorithm for our blockchain, implementation of our cryptography tools, and our blockchain management service.

Continuing with Phase 3, we will then implement our admin view, user management service, company registration service, new block creation process, and policy management service. Lastly, with Phase 4, we will then focus on UI polishing, final testing, and deployment. The timescale for each of these phases is approximately a month, keeping us on track with our end deadline with an additional slight buffer if needed.

1. “GDPR Key Changes”, https://eugdpr.org/the-regulation/ [↑](#footnote-ref-0)