Thesis Prospectus

A Thesis Prospectus submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Johnathan Middleton

Spring, 2024

Thesis Prospectus

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

William Davis, Department of Science, Technology and Society

Thesis Prospectus

Forgetting is necessary for humans. While we have kept historical records for years, and our minds are a form of these, rooted in all of these is imperfection. Records are misplaced and memories are fragile and easily lost. Without this imperfection, we would not be able to make social progress. As an individual, we forget things all the time (Kleiner et al. 69). Think about the last argument you had. It's likely it didn't cease to bother you because a perfect resolution was reached, but rather because it simply floated out of your headspace. The same is true of society. We must forget, collectively, the good and the bad of our history in order to continue to progress (Bucerius et al. 534). While we have laid a solid path forward in the fast-moving world of information towards data collection and aggregation, we have not started to challenge our goals and motivations in the matter.

Public discourse and legislature lags behind on the matters of data collection (Kovanič and Spáč 187) and in order to steer the ship of information security in the right direction, we need to reconsider our ideals when it comes to the collection of individual information. Most notably in the public sphere of legislation there exists the General Data Protection Regulations, hereafter referred to as GDPR (EU, 2016). While their inception was groundbreaking and deeply necessary, it will not come as a surprise to hear that this internationally relevant legislation pertaining to rapidly evolving topics in the technology industry has not been able to keep up. Additionally, the ideals brought forward by the GDPR are not necessarily in line with those of our societal norms, as social waves are brought on by platform after platform. Pauliina Hirvonen and Kari (576) discuss in addition the lack of "situational awareness" the points addressed in the GDPR have in the workforce. While personal values of privacy and information security remain strong, there is a strong disconnect between them and the way they are practiced in the workforce. In order to close this gap, we must create a durable framework that clearly outlines both the interactions between individuals and these technologies, as well as explicitly declares boundaries between the two.

A Data Pipeline in the Technological Industry

Within the technical aspect of this study, I bring forward a data pipeline orchestrated in a large technology corporation, namely Amazon. The data pipeline is responsible for collecting advertising data from users of Amazon's search services and aggregating this data in order to produce a business review sheet. A critical component of this process is "data provenance," which consists of the metadata attached to pieces of information that describe its origin and history (Pan et al. 1). This means that user advertising data can be collected over large periods of time and accurately back-tracked. The tools enabling this are relatively new when compared to the progress in data privacy, and which include cloud computing and machine learning algorithms. This, paired with advancements in data storage capabilities (Maha Dessokey et al. 460) and collection methods, have led to a lapse in the synchronicity of individual ability of memory and its technological equivalents within the commercial industry. In a study done on a group of youths ages 10-21 in the United States, it was found that while the participants generally had a competent understanding of the nature and volume of data being collected on them, they were severely lacking in the perception of data retention (Goray and Schoenebeck 25). Through comparative analysis of these results and terms of service agreements, a large disparity was found in this understanding. The study was performed in the context of social media companies, which are infamous for their methods of both passive and active data collection - passive collection being data created automatically, often unbeknownst to the user, in contrast with active collection in which the data is actively created by the user. However, these processes are not unique to the social media industry. Digital advertising has evolved to an entire field in itself, proving itself an enormously profitable strategy that uses the internet to provide highly targeted, instantaneous advertising. The "digital footprint" is a common term used to describe the path of personal and usage information of users of online applications. The information involved in this resource is highly coveted by data collectors of any tract (Cinar and Ateş, 10) as it provides a medium by which one can predict and influence consumer behavior.

Consequences of the Digital Footprint

The data pipeline previously mentioned is one of many examples of the process of this collection of the digital footprint. Data governance is the proposed solution with increasingly growing popularity. Concerns about consumer trust and data privacy are growing, and in recent years the COVID-19 pandemic has changed the landscape, as numerous governments began using unprecedented methods of data collection (Zarouali et al. 2049). Trust in these processes, when data leaks, large scale hacking attempts, and breaches of privacy seem to be all too common, is critical for them to remain functioning. Data governance describes a framework in which the roles, responsibilities, and processes of the layers of data collection are clearly defined. Within the private industry, where existing regulations are often bent and abused, there needs to be a solid framework that limits in particular the longevity of data that exists within the massive servers of dominant industry leaders, and which holds the continuously collected personal and private information of millions of users.

Societal Outlook on Data Permanence and Collection

Through this lens of data pipeline aggregation, I focus on the contrast between the perception of data collection and permanence and the actuality of its longevity and retention in systems throughout the technological sector. The onset of data collection in the modern technological industry has left a significant gap in the way our brains work and how we utilize records. The simple act of forgetting plays a critical role in our cognition (Bayliss and Jarrold 163) and this shortcoming, if you want to call it that, affects our history, our society, how we perceive our leaders, our ancestors, and much more (Schwartz 478). It is further proclaimed that this act of forgetting is necessary for social order and progress (Rieff 7). By failing to account for this disconnect between what Rieff might argue is a fundamental trait of one's humanity and the complete lack thereof in the modern world, we open ourselves up to the possibility of severe social consequences. Rieff's work is highly controversial, and Schwartz highlights an example where historical records led to the imprisonment of a guilty party. However, in the current context, we are not comparing a good memory with a better one. Instead, in the age of information, we have access to a "perfect" memory, which can be filed and sorted and accessed in more ways than we can imagine.

The Perfect Memory of Technology

We have leaped past the point of human capability when it comes to memory, and are in a position where the act of forgetting has been long lost. Highly superior autobiographical memory (HSAM) is a relatively newly researched phenomenon in which the subject is able to recall large amounts of autobiographical information (LePort et al. 78). While the emotional effects of this condition have not been deeply researched, from preliminary results it becomes apparent that these individuals who are able to recall highly specific memories from their life seemingly at ease run into problems with letting go of grudges and moving on from traumatic events. This is a localized example of the contrast, however it can be extrapolated to the wider scaled issues that we face today. To relate this to the data collection processes described previously, we now see social media companies performing a methodical retention of activity, interaction, and much more (Yang et al., 2022). The social landscape has been transformed by the popularization of these technologies, to the extent that many interactions can be defined by them, which in and of itself is not necessarily a positive consequence of their adoption (Cho et al. 1). In turn, we have effectively replicated a HSAM within our social networks.

Consequences of a Perfect Memory

There already exist consequences of this paradigm, including a growing trend of social ostracism for the exercise of free speech shared online (Hu and Barradas 609), where the permanence of social media harms those who use it. Hu and Barradas (610) illustrate the dramatic and permanent consequences of conflicts arising on these platforms and how it results in a phenomenon of self-censorship, in which individuals voluntarily withhold personal opinions or information in order to avoid what might come with it. As individuals with imperfect memories, we are designed to forget and we conduct our social interactions and relationships as such. Article 17 of the aforementioned GDPR is the “Right to erasure”, aptly subtitled the ‘right to be forgotten’ (EU, 2016). Within it are guidelines detailing the unlawful nature of permanent data collection in certain circumstances, and while it is dated, it shows the beginnings of effective legislation on the matter. As individuals we carry these values of forgetting innately, as Rieff would say, and yet we are lagging behind in terms of our deliverance on the issue. With domineering technology shaping our social lives and hurting those in the process, it’s necessary to prepare ourselves for the unpredictable future consequences. Evaluating this situation in the context of data governance gives us an idea of what a solution might look like. Enacting a framework in which the longevity of data is monitored and erased under realistic guidelines eliminates the brutal permanence in which interactions are conducted on the internet. Especially in the highly personal world of social media, we can see just how dangerous the lack of these barriers is.

Conclusion

With the analysis of a data pipeline used in the advertising industry currently coupled with the study of the consequences of "forgetting to forget," I aim to highlight a necessary change in goal and motivation for the data collection industry that will promote more socially beneficial practices within the age of information. Additionally through promotion of the extension of existing GDPR legislation, I hope to express how these problems and their solutions manifest in the industry. Oftentimes it's difficult to envision a piece of legislature enacting itself in the real world, especially in the case of data privacy or technology in general. The age of information reigns supreme and the constant stream has integrated itself deeply into our daily lives. It can be difficult to look through this and see how we might be negatively affected by practices at hand. Ideally through the portrait of a data aggregation pipeline that is commonly used throughout the industry we can see the intimacy data privacy has with each and every one of us, even in these large corporations, which are seemingly disconnected from our private lives at their size.

References

Bayliss, D. M., & Jarrold, C. (2015). How quickly they forget: The relationship between forgetting and working memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 163–177. https://doi.org/10.1037/a0037429

Bucerius, S. M., Thompson, S. K., & Dunford, D. T. (2022). Collective Memory and Collective Forgetting: A Comparative Analysis of Second-Generation Somali and Tamil Immigrants and Their Stance on Homeland Politics and Conflict. *Qualitative Sociology*, 553–556. https://doi.org/10.1007/s11133-022-09508-4

Cho, H., Li, P., Ngien, A., Marion Grace Tan, Chen, A., & Elmie Nekmat. (2023). The bright and dark sides of social media use during COVID-19 lockdown: Contrasting social media effects through social liability vs. social support. *Computers in Human Behavior*, 1–11. https://doi.org/10.1016/j.chb.2023.107795

Cinar, N., & Ateş, S. (2022). Data Privacy in Digital Advertising: Towards a Post Third-Party Cookie Era. *SSRN Electronic Journal*, 1–27. https://doi.org/10.2139/ssrn.4041963

EU. (2016, April 27). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

Goray, C., & Schoenebeck, S. (2022). Youths' Perceptions of Data Collection in Online Advertising and Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27. https://doi.org/10.1145/3555576

Hu, W., & Barradas, D. (2023). Work in Progress: A Glance at Social Media Self-Censorship in North America. *IEEE Xplore*, 609–618. https://doi.org/10.1109/EuroSPW59978.2023.00072

Kleiner, M. M., Nickelsburg, J., & Pilarski, A. M. (2012). Organizational and Individual Learning and Forgetting. *ILR Review*, *65*(1), 68–81. https://doi.org/10.1177/001979391206500104

Kovanič, M., & Spáč, S. (2022). Conceptions of Privacy in the Digital Era: Perceptions of Slovak Citizens. *Surveillance & Society*, *20*(2), 186–201. https://doi.org/10.24908/ss.v20i2.14099

LePort, A. K. R., Mattfeld, A. T., Dickinson-Anson, H., Fallon, J. H., Stark, C. E. L., Kruggel, F., Cahill, L., & McGaugh, J. L. (2012). Behavioral and neuroanatomical investigation of Highly Superior Autobiographical Memory (HSAM). *Neurobiology of Learning and Memory*, *98*(1), 78–92. https://doi.org/10.1016/j.nlm.2012.05.002

Maha Dessokey, Saif, S. M., Hesham Eldeeb, Salem, S. A., & Saad, E. M. (2022). Importance of Memory Management Layer in Big Data Architecture. *International Journal of Advanced Computer Science and Applications*, *13*(5), 460–466. https://doi.org/10.14569/ijacsa.2022.0130554

Pan, B., Stakhanova, N., & Ray, S. (2023). Data Provenance in Security and Privacy. *ACM Computing Surveys*, *55*(14s), 1–35. https://doi.org/10.1145/3593294

Pauliina Hirvonen, & Kari, M. J. (2023). Building Situational Awareness of GDPR. *European Conference on Cyber Warfare and Security*, *22*(1), 575–583. https://doi.org/10.34190/eccws.22.1.1077

Rieff, D. (2016). *In praise of forgetting : historical memory and its ironies* (pp. 1–145). Yale University Press.

Schwartz, B. (2017). The Future of Memory. *Society*, *54*(5), 478–484. https://doi.org/10.1007/s12115-017-0176-z

Yang, J., Xiu, P., Sun, L., Ying, L., & Muthu, B. (2022). Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management*, *59*(1). https://doi.org/10.1016/j.ipm.2021.102751

Zarouali, B., Strycharz, J., Helberger, N., & de Vreese, C. (2022). Exploring people's perceptions and support of data-driven technology in times of COVID-19: the role of trust, risk, and privacy concerns. *Behaviour & Information Technology*, 2049–2060. https://doi.org/10.1080/0144929x.2021.2022208