THE UNIVERSITY OF CHICAGO

TOWARDS RECLAIMING DATA AGENCY IN THE AGE OF UBIQUITOUS MACHINE LEARNING

A DISSERTATION PROPOSAL SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

BY
EMILY WENGER

CHICAGO, ILLINOIS
JUNE 2022

TABLE OF CONTENTS

ABSTRACT iii

1 INTRODUCTION 1

2 PROPOSED THESIS CONTENT 4
2.1 Data Agency via Disruption 4
2.2 Data Agency via Tracing 6
2.3 Data Agency via Direct Attack 7
2.4 A Framework for Understanding Data Agency Solutions 9
2.5 The Future of Data Agency 10
2.5.1 Mapping Design Space 10
2.5.2 Open Challenges 11

3 DISSERTATION TIMELINE 12

REFERENCES 14

ABSTRACT

As machine learning (ML) models have expanded in size and scope in recent years, so has the amount of data needed to train them. This creates privacy risks for individuals whose data – be it their images, emails, tweets, or browsing history – is used for training. For example, ML models can memorize their training data, revealing private information about individuals in the dataset [17, 6]. Furthermore, users whose data is co-opted for ML use may end up enrolled in a privacy-compromising system, such as a large-scale facial recognition model [23, 2]. Most existing work on ML data privacy accepts the premise that data use is inevitable and instead tries to mitigate privacy risks during model training [15, 35, 19]. However, privacy-conscious individuals may desire agency over *how and if* their data is used, rather than only having their privacy preserved when it is used. *Data agency*, the ability to know and control how and if one's data is used in ML systems, is an important complement to existing privacy protection approaches, and it is the focus of this thesis.

Data agency can take many forms, and this thesis will develop *technical solutions that enable individuals to disrupt or discover when their data is used in large-scale ML systems.* It targets data agency in the context of large-scale facial recognition (FR) systems, providing ways for users to combat unwanted facial recognition. This work proposes three data agency solutions to disrupt or trace data use in FR systems or, in extreme cases, directly attack the FR system. Additionally, it develops a framework for reasoning about broadly about data agency in the context of FR. It will use this framework to outline both technical and social challenges of data agency solutions in the FR space and propose directions for future research. Finally, this thesis will discuss the connections between the proposed FR-specific data agency solutions and methods for reclaiming data agency in other domains.

CHAPTER 1

INTRODUCTION

In recent years, machine learning (ML) models have been eagerly adopted to perform a variety of tasks, from real-time language translation to facial recognition. As machine learning models have expanded in use, size, and scope, so has the amount of data needed to train them. To feed their data-hungry models, ML practitioners use a variety of data sources, from public websites [23, 46] to private application data [10] to surveillance data from public spaces [48]. While sometimes this data is obtained with consent and user knowledge (e.g. a large scale user study), there are numerous well-documented cases of user data being co-opted for ML use without user consent. For example, the facial recognition company Clearview.ai has developed a facial recognition model using over 3 billion images scraped from social media sites [23]. Relatedly, the AI research company DeepMind has faced a lawsuit for secretly funneling data from a UK health services app to train ML models for medical diagnosis [3].

These examples imply that data from individuals who use an online service or occupy a public space may be used to train ML models. Beyond issues of consent, such data use poses a serious risk to individual privacy. Prior work has shown that models can memorize and regurgitate their training data, revealing private information about individuals whose data is in the dataset [17, 6]. Furthermore, users whose data is co-opted for ML use may end up enrolled in a privacy-compromising system, such as a large-scale facial recognition model licensed to whoever wants to pay [23].

Numerous solutions have been proposed to address the issue of data privacy during ML training, each of resting on a different understanding of privacy. The first set of solutions take the perspective that "only the user should know the exact content of their data," inspiring privacy solutions like federated learning and encrypted training [35, 19]. These solutions use locally hosted data or encryption methods to limit third party access to the data. Another

set of solutions simply requires that "data should not be traceable to the user who created it." This inspires solutions like differential privacy [15], which ensures that any single data point has no outsized influence model training, limiting an attacker's ability to reverse-engineer it.

While these solutions are effective in their own way, they share a common limitation: they assume data use is inevitable and merely mitigate privacy risks during model training and inference, after data has already been taken. Though important, post-facto privacy protection is not the only way to address these issues. Privacy-conscious individuals may desire agency over *how and if* their data is used, rather than only having their privacy preserved when it is used. *Data agency*, the ability to know and control how and if one's data is used in ML systems, is an important complement to existing privacy protection approaches, and it is the focus of this thesis.

At a high level, restoring data agency to individuals involves creating technical solutions that enable individuals to disrupt or discover when their data is used in large-scale ML systems. These technical solutions can and should be augmented by organizational and legislative efforts, but can often act quicker than these and so are the first line of defense against unwanted data use. The most straightforward approach to data agency is to *disrupt model training* by rendering data taken without user consent unusable in an ML setting. In the event where disruption is infeasible (e.g. the user has too little knowledge of the downstream ML setting in which it is used), a second data agency solution would be to *detect if the data is used to train a model.* This would allow users to know which of their data have been used without their consent, perhaps enabling future legal action. Finally, in extreme cases, data agency can involve *attacking illicit ML systems.* Depending on the user's goals, these attacks could either provide protection for users by embedding controlled misclassification behavior in the model or could be used to draw public attention the problem of data being used without consent.

Thesis Contribution

This thesis will both define the problem of ensuring data agency against ubiquitous ML systems and develop several technical solutions to address it. This thesis proposes three solutions, each of which use a different method of reclaiming data agency: **disruption** (§2.1), **tracing** (§2.2), and **direct attacks** (§2.3). For concreteness, the tools developed target data agency in the context of facial recognition, a machine learning technology that has recently become a flashpoint for civil liberties and privacy issues. Leveraging these tools and other insights, it then will **develop a framework** (§2.4) for thinking broadly about data agency against unwanted FR systems. This framework can serve as a template for the development of similar data agency frameworks in other domains. Finally, it will provide a **forward-looking analysis** (§2.5) of the viability of data agency techniques in the context of FR and beyond.

CHAPTER 2

PROPOSED THESIS CONTENT

2.1 Data Agency via Disruption

The first tool proposed is a poisoning attack against large-scale FR systems, which disrupts FR by making face images unusable. As alluded to previously, today's proliferation of powerful facial recognition models poses a real threat to personal privacy. Kashmir Hill from the New York Times recently reported on *Clearview.ai*, a private company that collected more than 3 billion online photos and trained a massive model capable of recognizing millions of citizens, all without knowledge or consent [1].

Opportunities for misuse of this technology are numerous and potentially disastrous. Anywhere we go, we can be identified at any time through street cameras, video doorbells, security cameras, and personal cellphones. Stalkers can find out our identity and social media profiles with a single snapshot [45]. Stores can associate our precise in-store shopping behavior with online ads and browsing profiles [34]. Identity thieves can easily identify (and perhaps gain access to) our personal accounts [14].

We believe that private citizens need tools to reclaim data agency against unwanted FR and protect themselves from being identified by unauthorized FR models. Unfortunately, previous work in this space is sparse and limited in both practicality and efficacy. Some have proposed distorting images to make them unrecognizable and thus avoiding facial recognition [56, 29, 47]. Others produce adversarial patches in the form of bright patterns printed on sweatshirts or signs, which prevent facial recognition algorithms from even registering their wearer as a person [57, 49]. Finally, given access to an image classification model, "clean-label poison attacks" can cause the model to misidentify a single, preselected image [43, 61].

Instead, we will propose a system to help individuals to inoculate their images against unauthorized facial recognition models at any time without significantly distorting their

own photos, or wearing conspicuous patches. Our system will achieve this by helping users adding imperceptible pixel-level changes ("cloaks") to their own photos. For example, a user who wants to share content (*e.g.* photos) on social media or the public web can add small, imperceptible alterations to their photos before uploading them. If collected by a third-party "tracker" and used to train a facial recognition model to recognize the user, these "cloaked" images would produce functional models that consistently misidentify them.

Our disruption or "cloaking" algorithm will take the user's photos and compute minimal perturbations that shift them significantly in the feature space of a facial recognition model (using real or synthetic images of a third party as a landmark). Any facial recognition model trained using these images of user will learn an altered set of "features" of what makes them look like them. When presented with a clean, uncloaked image of the user, *e.g.* photos from a camera phone or streetlight camera, the model will find no labels associated with the user in the feature space near the image, and classifies the photo to another label (identity) nearby in the feature space. Our proposal will explore several key dimensions of data agency via data disruption:

- We will explore how to produce significant alterations to images' feature space representations using perturbations imperceptible to the naked eye.
- We will evaluate how our system performs regardless when the tracker uses different techniques to train their model (e.g. transfer learning or from scratch).
- We will test against state-of-the-art facial recognition services from Microsoft (Azure Face API), Amazon (Rekognition), and Face++.
- In challenging scenarios where clean, uncloaked images are "leaked" to the tracker and used for training, we will show explore whether a *Sybil* identity can boost privacy protection.
- Finally, we will consider a tracker who is aware of our image cloaking techniques and evaluate the efficacy of potential countermeasures.

2.2 Data Agency via Tracing

The next data agency tool takes a different approach, again in the context of FR. Instead of trying to fully disrupt a FR system's operation, it will instead provide means for individuals to determine whether their images have been used to train a model. As prior work has discussed [42], preventing FR completely is difficult, due to the power imbalance between FR system operators and individuals. However, tracing data through a FR system may be more tractable. Data privacy laws grow more strict, having the ability to trace data through models may provide individuals leverage to pursue legal action against unwanted FR. Data agency through tracing endows individuals with the ability to *prove* data misuse, which may ultimately be more powerful than data disruption.

To this end, we will propose a *datamarking* scheme, in which users subtly "tag" images they post online. If stolen and used for model training, the tagged images can be detected through their influence on the model's prediction behaviors. We will demonstrate the efficacy of this scheme in several large-scale image dataset and in real-world scenarios.

Our proposal will include the following evaluations:

- We will develop a set of empirical properties a viable datamark must satisfy.
- We will proposal a novel datamarking scheme and demonstrate the efficacy of that scheme for both facial and object classification tasks.
- We will verify that datamarking remains effective even when multiple datamarks are present within the model or within a single class.
- We will explore how different conditions (e.g. training decisions, model and dataset size, etc.) affect datamark detectability.
- We will test our datamarking scheme against real-world facial recognition services such as Amazon Rekognition and Face++.
- We will consider datamark disruption strategies such as removal and detection and ensure

that our datamark persists against adapative adversaries who wish to remove it.

2.3 Data Agency via Direct Attack

Finally, we propose a tool which reclaims data agency via a direct attack on a FR system. Despite their known impact on numerous applications from facial recognition to self-driving cars, deep neural networks (DNNs) are vulnerable to a range of adversarial attacks [7, 40, 27, 39, 32, 4, 11]. One such attack is the backdoor attack [20, 33], in which an attacker corrupts (*i.e.* poisons) a dataset to embed hidden malicious behaviors into models trained on this dataset. These behaviors only activate on inputs containing a specific "trigger" pattern.

Backdoor attacks are sneaky because corrupted models operate normally on benign inputs (*i.e.* achieve high classification accuracy), but consistently misclassify any inputs containing the backdoor trigger. This property has galvanized efforts to investigate backdoor attacks and their defenses, from government funding initiatives (*e.g.* [51]) to numerous defenses that either identify corrupted models or detect inputs containing triggers [9, 18, 21, 41, 54]. Critically, backdoor attacks against facial recognition could be used to enable data agency, as they enable users to cause controlled misclassifications in FR systems. These misclassifications can both serve as indicators that models are trained on stolen data (reminiscent of the prior data tracing solution) and allow individuals fight back by corrupting models trained illicitly on their data.

Current literature on backdoor attacks and defenses mainly focuses on *digital* attacks, where the backdoor trigger is a digital pattern (*e.g.* a random pixel block) that is digitally inserted into an input. These digital attacks assume attackers have run-time access to the image processing pipeline to digitally modify inputs [26]. This rather strong assumption significantly limits the applicability of backdoor attacks to real-world settings. This proposal will consider a more realistic form of the backdoor attack against facial recognition systems. It will use everyday, physical objects as backdoor triggers, included naturally in training

images, thus eliminating the need to compromise the image processing pipeline to add the trigger to inputs. An attacker could then activate the attack simply by wearing/holding the physical trigger object, e.g. a scarf or earrings. We call these "physical" backdoor attacks.

To evaulate physical backdoor performance, we will perform a detailed empirical study on the training and execution of physical backdoor attacks under a variety of real-world settings. We will focus primarily on the task of facial recognition due to the key application of physical backdoor attacks as AFR tools. Using 7 physical objects as triggers, we will collect an IRB-approved custom dataset. We will then launch backdoor attacks against three common face recognition models (VGG16, ResNet50, DenseNet) by poisoning their training dataset with our image dataset. We will adopt the common (and realistic) threat model [20, 31, 30, 52], where the attacker can corrupt training data but cannot control the training process.

Our evaluation will perform several key analyses:

- We will use the BadNets method [20] to generate backdoored models and evaluate whether physical backdoors are effective.
- We will explore different attack properties and threat model assumptions to isolate key factors in the effectiveness of physical backdoor attacks.
- We will relax our threat model and explore whether attackers can still succeed when constrained to poisoning a small fraction of classes in the dataset. Additionally, we will evaluate whether models poisoned by backdoors based on digitally injected physical triggers can be activated by a subject wearing the actual physical triggers at run-time.
- Finally, we will study the effect of physical backdoors on state-of-the-art backdoor defenses.

2.4 A Framework for Understanding Data Agency Solutions

Next, this thesis will propose and use a *stage-based* framework to categorize current data agency solutions targeting unwanted FR. These solutions, referred to as "anti-facial recognition" or AFR tools, provide a representative sample of methods for reclaiming data agency. Although this framework is designed for AFR specifically, both it and the method of its construction present generalizable principles for reasoning about data agency in other contexts.

We choose to systemize AFR tools as a "model" for thinking about data agency methods because they have attracted significant attention from the research community. In the last 12 months, more than a dozen AFR tools have been proposed (e.g., [28, 55, 44, 24, 16, 12, 13, 58, 50, 59, 8, 60, 37, 25, 5, 53]). While most are constrained to research prototypes, a few of these tools have produced public software releases and gained significant media attention [44, 12, 22].

Proposals in the rapidly growing collection of AFR tools differ widely in their assumptions and techniques, and target different pieces of the facial recognition pipeline. There is a need to better understand their commonalities, to highlight performance tradeoffs, and to identify unexplored areas for future development. Existing surveys [36, 38] on facial privacy issues do not address AFR tools. [36] summarizes proposals that modify faces (and other biometric data) to allow automatic extraction of a user's identity but not other sensitive information, an orthogonal goal to that of AFR. [38] provides a comprehensive survey on facial privacy protection methods but was published before the rise of DNN-based FRs and AFRs.

This portion of the thesis will address this need by providing a common framework for analyzing a wide range of AFR systems. More specifically, it will make the following contributions:

- **Taxonomy of targets in facial recognition systems:** AFR systems target a wide range of components in the facial recognition process. Using a generalized version of the facial recognition data pipeline, we will provide a framework to reason broadly about

existing and future AFR work.

- **Categorization and analysis of AFR tools:** We will take the current body of work on AFR tools, and categorize and analyze them using our proposed framework.
- **Mapping design space based on desired properties:** We will identify a core set of key properties that future AFR systems might optimize for in their design, and provide a design roadmap by discussing how and if such properties can be achieved by AFR systems that target each stage in our design framework.
- **Open challenges:** We use our framework to identify significant challenges facing current AFR systems, as well as directions for potential solutions.

2.5 The Future of Data Agency

The final thesis component will outline a vision for future work on data agency. Again using FR and AFR tools as a starting point, it will do so through two mechanisms: a *mapping of the data agency design space* and *a discussion of open challenges.*

2.5.1 Mapping Design Space

To provide a method of reasoning about optimal data agency solutions, we will use a case study in FR. First, we will consider the design space of AFR tools through the lens of specific FR stages they disrupt. To date, all existing AFR proposals have focused their design around disrupting a single stage in this framework. Assuming an AFR tool must disrupt some portion of the FR pipeline to be effective, we will map out and explore the design space of AFR tools using this framework.

For researchers and practitioners in the AFR community, perhaps the most critical question is: *"what are the benefits and limitations of AFR tools that target each specific framework stage?"* Or, an alternative form of the question might be: *"Given a set of prioritized prop-*

erties for an AFR system, can I find the best stage(s) to disrupt in order to achieve them?"

We will attempt to answer these questions here, by first identifying a set of high level properties that AFR tools can potentially optimize for, then for each property, discussing how targeting a given stage affects an AFR tool's ability to achieve it. Ultimately, we hope to provide a high level roadmap that can guide the design of AFR tools optimizing for specific properties in mind. While we consider each stage in isolation, it might be possible for an AFR tool to target multiple stages, gaining a combination of benefits (and limitations).

After discussing these concepts through the lens of AFR, we will consider how they might generalized to data agency in other domains.

2.5.2 Open Challenges

Finally, we will describe what we see as the major technical and broader social/ethical challenges facing future data agency tool development, first in the AFR space and then more broadly. Each challenge will multiple properties and stages discussed throughout this thesis. For each challenge, we will provide context for why the challenge exists and, where possible, suggest ways to address it. The challenges described will represent our best efforts to understand and systematize the space. They will not be exhaustive, and are meant as signposts rather than a comprehensive roadmap.

CHAPTER 3

DISSERTATION TIMELINE

Summer 2022

June 17: PhD Proposal/Advance to Candidacy

July - August: Complete systemization of the anti-facial recognition space.

July 12-14: Attend Veritas Forum Young Faculty summit.

July 17-19: Attend International Conference of Machine Learning (ICML) conference in Baltimore.

July - August: Finish project on data provenance tracking in facial recognition models.

August 10-12: Attend USENIX Security Symposium in Boston for presentation of "Blacklight" paper.

Fall 2022

August 19: Submit data provenance tracking paper to 2023 USENIX Security Symposium.

August - December: Develop faculty job application materials: research statement, teaching statement, job talk.

October 1: Target completion date for systemization of anti-facial recognition space.

November - December: Start thesis writing.

November 1: Submit SALSA paper to International Conference on Learning Representations (ICLR).

November 29 - December 1: Attend the Conference on Neural Information Processing Systems (NeurIPS) in New Orleans.

December 15: Submit faculty job applications.

Winter 2023

January - March: Travel to faculty job interviews.

January - April: Continue thesis writing.

Spring 2023

April - May: Prepare for thesis defense.

Early May: Thesis defense.

June 3: Commencement ceremony.

REFERENCES

[1] Clearview.ai.

[2] Pimeyes.

[3] Deepmind faces legal action over nhs data use, 2021.

[4] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proc. of ICLR*, 2018.

[5] Kieran Browne, Ben Swift, and Terhi Nurmikko-Fuller. Camera adversaria. In *Proc. of CHI*, pages 1–9, 2020.

[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *Proc. of USENIX Security*, 2021.

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. of IEEE S&P*, 2017.

[8] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. Face-off: Adversarial face obfuscation. *Proc. of PETS*, 2021.

[9] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and B̃iplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

[10] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2287–2295, 2019.

[11] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proc. of AISec*, 2017.

[12] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *Proc. of ICLR*, 2021.

[13] Thomas Cilloni, Wei Wang, Charles Walter, and Charles Fleming. Ulixes: Facial recognition privacy with adversarial machine learning. *Proc. of PETs*, 2022.

[14] Jim Cross. Valley attorney: Facebook facial recognition carries identity theft risk. *KTAR News*, September 2019.

[15] Cynthia Dwork. Differential privacy: A survey of results. In *Proc. of International Conference on Theory and Applications of Models of Computation*. Springer, 2008.

[16] Ivan Evtimov, Pascal Sturmfels, and Tadayoshi Kohno. Foggysight: a scheme for facial lookup privacy. *Proc. of PETS*, 2021.

[17] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. of CCS*, 2015.

[18] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proc. of ACSAC*, 2019.

[19] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *Proc. of ICML*, 2016.

[20] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[21] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in AI systems. *arXiv preprint arXiv:1908.01763*, 2019.

[22] Adam Harvey. Cv dazzle: Camouflage from face detection. *Master's thesis*, 2010.

[23] Rebecca Heilweil. The world's scariest facial recognition company, explained. *Vox.com*.

[24] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *Proc. of ICLR*, 2021.

[25] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *Proc. of ICPR*. IEEE, 2021.

[26] Ram Shankar Siva Kumar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xiã. Adversarial machine learning–industry perspectives. *arXiv preprint arXiv:2002.05646*, 2020.

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proc. of ICLR*, 2017.

[28] Tao Li and Min Soo Choi. Deepblur: A simple and effective method for natural image obfuscation. *arXiv preprint arXiv:2104.02655*, 1, 2021.

[29] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proc. of CVPR*, 2019.

[30] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020.

[31] Cong Liao, Haoti Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *arXiv preprint arXiv:1808.10307*, 2018.

[32] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. of ICLR*, 2016.

[33] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proc. of NDSS*, 2018.

[34] Angelica Mari. Brazilian retailer quizzed over facial recognition tech. *ZDNet*, March 2019.

[35] H Brendan McMahan, Eider Moore, Daniel Ramage, S Hampson, and B Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[36] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Tran. on Information Forensics and Security*, 2021.

[37] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. Adversarial light projection attacks on face recognition systems: A feasibility study. In *Proc. of CVPR*, 2020.

[38] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015.

[39] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proc. of Asia CCS*, 2017.

[40] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Proc. of Euro S&P*, 2016.

[41] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. In *Proc. of NeurIPS*, 2019.

[42] Evani Radiya-Dixit and Florian Tramer. Data poisoning won't save you from facial recognition. *arXiv*, 2021.

[43] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Proc. of NeurIPS*, 2018.

[44] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proc. of USENIX Security*, 2020.

[45] Maya Shwayder. Clearview AI's facial-recognition app is a nightmare for stalking victims. *Digital Trends*, January 2020.

[46] Olivia Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. *NBC News*, 2019.

[47] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proc. of ECCV*, 2018.

[48] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):1–14, 2018.

[49] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proc. of CVPR (workshop)*, 2019.

[50] Marc Treu, Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Fashion-guided adversarial attack on person segmentation. In *Proc. of CVPR*, 2021.

[51] Trojans in artificial intelligence (TrojAI), Feb. 2019. `https://www.iarpa.gov/index.php/research-programs/trojai`.

[52] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[53] Nicholas Vincent and Brent Hecht. Can "conscious data contribution" help users to exert "data leverage" against technology companies? *Proc. of CHI*, 2021.

[54] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proc. of IEEE S&P*, 2019.

[55] Yunqian Wen, Li Song, Bo Liu, Ming Ding, and Rong Xie. Identitydp: Differential private identification protection for face images. *arXiv preprint arXiv:2103.01745*, 2021.

[56] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-Protective-GAN for face de-identification. *arXiv:1806.08906*, 2018.

[57] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. *arXiv:1910.14667*, 2019.

[58] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Proc. of ECCV*, 2020.

[59] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Proc. of ECCV*, 2020.

[60] Mingfu Xue, Shichang Sun, Zhiyu Wu, Can He, Jian Wang, and Weiqiang Liu. Socialguard: An adversarial example based privacy-preserving technique for social images. *arXiv preprint arXiv:2011.13560*, 2020.

[61] Chen Zhu, W Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *Proc. of ICML*, 2019.