

CARR CENTER FOR HUMAN RIGHTS POLICY HARVARD KENNEDY SCHOOL

The Ethical Use of Personal Data to Build AI Technologies:

A Case Study on Remote Biometric Identity Verification

Neal Cohen

Carr Center
Discussion Paper



The Ethical Use of Personal Data to Build Artificial Intelligence Technologies: A Case Study on Remote Biometric Identity Verification

Neal Cohen
Technology and Human Rights Fellow
Carr Center for Human Rights Policy
Harvard University

April 2020

Neal Cohen is the Director of Privacy at Onfido Limited, and this paper was written as part of the Technology and Human Rights Fellowship at Harvard's Kennedy School's Carr Center for Human Rights Policy. Neal is also a non-residential fellow at Stanford's Center for Internet & Society and previously a research fellow at Harvard's Berkman Klein Center for Internet & Society. Prior to joining Onfido Limited, Neal worked as an English and New York qualified lawyer at Perkins Cole LLP, White & Case LLP, and Dentons in the United Kingdom and the United States. This paper would not have been possible without the help of the research and engineering teams at Onfido. In particular, thank you to Martins Bruveris, Pouria Mortazavian, and Mohan Mahadevan for their continuous time and efforts in addressing the privacy challenges presented by artificial intelligence technologies. All thoughts within are that of the author and do not represent those of Onfido Limited or any other organization.

The Technology and Human Rights Fellowship is part of the Carr Center for Human Rights Policy's project to examine how technological advances over the next several decades will affect the future of human life, as well as the protections provided by the human rights framework.

Table of Contents

- 2 Introduction
- 3 Remote Biometric Identity Verification as an AI Technology
- 6 The Technical Requirements to Build an AI Technology to Perform Remote Biometric Identity Verification
- 9 The New Privacy Challenges of Using Personal Data to Build AI Technologies
- 11 Ethical Solutions to Using Personal Data to Build AI Technologies
- 14 Conclusion

ABSTRACT: Artificial Intelligence (AI) technologies have the capacity to do a great deal of good in the world, but whether they do so is not only dependent upon how we use those AI technologies but also how we build those AI technologies in the first place. The unfortunate truth is that personal data has become the bricks and mortar used to build many AI technologies and more must be done to protect and safeguard the humans whose personal data is being used. Through a case study on AI-powered remote biometric identity verification, this paper seeks to explore the technical requirements of building AI technologies with high volumes of personal data and the implications of such on our understanding of existing data protection frameworks. Ultimately, a path forward is proposed for ethically using personal data to build AI technologies.

Introduction

Throughout history, we have shaped the world around us through the use of technology. Early humans learned to build a hut using mud. Then, we learned to build houses using bricks and mortar, and more recently, we learned to build skyscrapers out of steel and glass with even more advanced technology.

As we have used technology to shape the world around us, so have we recognized the need to protect ourselves from this technology. For example, a skyscraper cannot be opened to the public without going through a health and safety review. Not only do we need to ensure that the building will not collapse, but we need to ensure the materials used are safe and will not harm us.¹

This is not only good practice, but such review is expected and required by society. We expect that when technology is used to build a product, that product will only be made available if it is safe. Laws exist to help control and mitigate the risks of technology, albeit these laws often trail behind the technological innovations they are designed to control.

The same expectation applies to products powered by Artificial Intelligence (AI) technologies.² However, AI technologies introduce a fundamental shift in how we build products. When building AI technologies, the bricks and mortar of the

past are now frequently replaced with the personal data of the wider public. As an example, building accurate and effective facial recognition technology requires millions upon millions of facial images to train the machine learning models used to power that technology. Without the use of those facial images belonging to real people, the technology cannot be built. The dependence on large amounts of personal data is common to AI technologies that are based on machine learning and deep neural networks.

This need to use personal data to build the underlying AI technology introduces a new challenge for how we build products. While we must ensure that AI technologies are safe for use when made available to the general public, we also must ensure the manner in which the technology is built is safe and does no harm.

Legal frameworks across the world are designed to ensure that digital technologies are safe for individuals. We see this in international agreements, data privacy laws, and consumer protection laws.³ Governments are also beginning to develop ethical frameworks for the application of AI.⁴ Yet, these laws and ethical frameworks tend to focus on the use of digital technologies. While they are often written broadly enough to cover the use of personal data to build AI technologies, it is not yet well understood how these laws and frameworks should be interpreted and applied when building AI technologies.

¹ As an example, many countries around the world regulate the use of asbestos in building materials.

² The Oxford Reference Dictionary defines “Artificial Intelligence” as “the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages”, available at <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960>. See also Future of Privacy Forum: The Privacy Expert’s Guide to Artificial Intelligence and Machine Learning (October 2018), available at https://fpf.org/wp-content/uploads/2018/10/FPF_Artificial-Intelligence_Digital.pdf

³ To name a few, such legislative frameworks include The Council of Europe Convention 108, available at <https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108>; European General Data Protection Regulation (REGULATION (EU) 2016/679) (“GDPR”), available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1465452422595&uri=CELEX:32016R0679>; the US Federal Trade Commission Act (15 U.S.C. §§ 41-58), available at <https://www.ftc.gov/enforcement/statutes/federal-trade-commission-act>; the California Consumer Privacy Act of 2018 (AB-375), available at https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375; the APEC Privacy Framework, available at [https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-\(2015\)](https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-(2015)); and the OECD Privacy Guidelines, available at <https://www.oecd.org/sti/ieconomy/privacy-guidelines.htm>.

⁴ See the European Commission’s Guidelines for Trustworthy AI, available at <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>; the International Conference of Data Protection and Privacy Commissioners Declaration on Ethics and Data Protection in Artificial Intelligence, available at https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf; and the White House’s Artificial Intelligence for the American People, available at <https://www.whitehouse.gov/ai/>.

This lack of clarity is particularly salient when an individual's personal data is used to build an AI technology. In such circumstances, there is no intended impact on that specific individual. Rather, the intended impact is on the individual who will be subject to the application of the AI technology.⁵ However, the very fact personal data is retained for the purpose of building the AI technology does introduce a risk that personal data can be lost, stolen, or otherwise used for an unintended purpose to the detriment of the individual—potentially significantly impacting the individual, dependent upon the specific personal data and how it is abused.

This is not to say that how AI technologies are used is any less important, but the intention of this paper is to shine a spotlight on how those AI technologies are built. There is still a critical need to closely examine how AI technologies are being used in society and to recognize the many harms and human rights violations that may be caused by them, especially where those AI technologies are used unlawfully or unethically.

Below, this paper seeks to explore how personal data is used when building AI technologies and how such personal data use can conform to the myriad of data privacy laws and ethical frameworks around the world. To help on this journey, a case study of AI-powered remote biometric identity verification (IDV) is used to demonstrate the challenges and potential solutions for using personal data to build AI technologies, specifically in regard to biometric facial verification technology. In exploring the issues highlighted through the building of IDV technology, this paper will discuss:

1. Remote biometric identity verification as an AI technology;
2. The technical requirements to build an AI technology to perform remote biometric identity verification;
3. The new privacy challenges of using personal data to build AI technologies; and
4. Ethical solutions to using personal data to build AI technologies.

Many of the thoughts below are based on the experience of building IDV technology with the research team at Onfido as well as Onfido's participation in the United Kingdom's Information Commissioner's Office (ICO) Privacy Sandbox.⁶

I. Remote Biometric Identity Verification as an AI Technology

In our daily lives, we transact with other people and companies throughout the day. Some of these transactions can and should be largely anonymous, while others require or necessitate that the transacting parties know each other's identity. For example, financial regulations require a bank to know the identity of an individual before giving that individual a bank account. This is necessary to guard against bad actors and money laundering.⁷ Another example, while not legally required, is two strangers meeting on a peer-to-peer service (such as a car sharing service, dating app, or online marketplace) who might feel safer if each person has identified themselves to the other person. This verified self-identification adds accountability and safety as the participating individuals know that if one of them breaks the law (for example, by robbing or otherwise harming the other person), there will be immediate consequences as there is a transactional record attached to their legal identity.⁸

In the past, identity verification typically took place in person and at the point of transaction. When opening a bank account, the customer would present their identity document to the bank teller, and the bank teller would assess the validity of the identity document and whether it belonged to the person presenting the identity document. Similarly, when renting a car, the driver would first go to the car rental company to present their driver's license to prove their identity and right to drive.

However, the information age forces us to consider a different paradigm for how to assert and prove identity without physical presence. Many online products and services have

⁵ To the extent an individual is later subjected to an AI technology after their personal data was used to build that AI technology, that individual will be directly impacted by that AI technology.

⁶ Within the ICO's Privacy Sandbox, Onfido is researching how to detect and mitigate algorithmic bias in machine learning models. The UK Centre for Data Ethics and Innovation (CDEI) is also observing this work for the purposes of their own work program on bias. For more details, see "AI-powered Onfido one of the first selected for the ICO's Sandbox" by Ali Vaziri, *Privacy Laws & Business* (UK Report Issue 105). Nothing in this paper should be understood or interpreted as an approval or endorsement from the ICO or the CDEI as to the statements within.

⁷ For an overview of many of the anti-money laundering laws and regimes around the world, see the information provided by the Financial Action Task Force (FATF) on its members and observers, available at <https://www.fatf-gafi.org/about/membersandobservers/>.

⁸ See *Business Insider*: "Uber's lax ID requirements in Brazil led to people playing 'Uber roulette' and sticking up drivers — leading to 16 murders", (August 23, 2019), <https://www.businessinsider.com/uber-roulette-driver-murders-brazil-super-pumped-book-2019-8>.

no physical premises through which an individual can have their identity verified, or are opting for a method of identity verification that is less cumbersome and more consistent in application, particularly in regard to fraud prevention.

In recent years, the need for remote identity verification has largely been filled by credit reference agencies that enable individuals to prove their identity by answering a series of knowledge-based questions about their specific identity. Yet, these checks are only effective to the extent that a person can be found in the credit reference agency's database, and unfortunately much of the world's population is not included—greatly impacting those excluded populations from participating in different aspects of society.⁹ Remote biometric identity verification seeks to overcome this barrier to inclusion by enabling any person with a government-issued identity document to quickly and accurately prove their identity without having to be present in a specific location or be included in any specific database.

IDV technology enables a person to prove their identity by processing information readily available to most individuals—images of their identity document and face. This technology uses machine learning models supplemented by human review experts, where needed, to determine the likelihood that (i) the identity document is genuine and not fraudulent; and (ii) the facial image on the identity document matches the provided image of the individual's face and is not a spoofed or fraudulent image.¹⁰ From this process, a detailed report is generated indicating the likelihood that the identity document is genuine and the two faces match, along with detailed reasons for such findings.¹¹

In most circumstances, IDV technology is made available to an individual as an integrated part of another service where the IDV technology is used for account opening and onboarding. While the individual has a direct relationship

with the company whose services the individual is trying to access, the individual's personal data will be processed by the IDV provider to carry out the IDV check. This is a common supply chain for AI technologies, but the data flow may not be apparent to the individual unless heightened notice is provided.

The onus is also on the company that is using the IDV technology to interpret and act upon the aforementioned report. This requires that company to understand what that report is saying about an individual and be able to recognize when an individual is having difficulty using the technology and requires additional help or recourse.¹² If that company is unable to responsibly and ethically use the IDV technology, there is a risk that the individual will be unable to access the requested service. This risk is especially concerning and significant where IDV is used as a barrier to participate in civil society or access vital services such as healthcare or banking.

As a real-life example of IDV technology, imagine an individual has just downloaded the mobile application for a bank that has no physical presence (such as a branch or a store). All banking services are provided entirely through the mobile application. Upon opening the mobile application, the individual is asked to create an account and go through an identity verification process so the bank may comply with their customer due diligence and anti-money laundering obligations. This process begins with the bank verifying the identity of the individual by asking that person to upload an image of their identity document and a live picture of their face. Through the use of IDV technology, the bank proves the legal identity of the individual by validating the individual's identity document and verifying that the live picture of the individual matches the facial image on the identity document.

Even though this paper uses IDV technology as a case study, it is also important not to lose sight of the other use cases for facial recognition technology and the harms and

⁹ See the World Bank, *Global Financial Development Report 2013 - Rethinking the Role of the State in Finance* (p. 134), available at <https://openknowledge.worldbank.org/handle/10986/11848>.

¹⁰ A spoofed or fraudulent image might be a person holding a photo of another person, a person wearing a mask, a person using face morphing technology, and other similar methods used to digitally impersonate another person.

¹¹ The report is not merely a single written document that is intended to be read by a person. While the report can be presented as such, each of the different findings within the report is also made available as an API call to the company using the IDV technology so that company can configure their own risk tolerances and decide how to respond to a given report (see footnote 31 for a definition of API call).

¹² While the company building the IDV must ensure that it is accessible for all types of individuals, the IDV provider likely does not have sight as to where specific individuals are not able to successfully interact with the IDV technology in a given moment.

There is still a critical need to closely examine how AI technologies are being used in society, and to recognize the many harms and human rights violations that may be caused by them.

human rights violations which might arise.¹³ This concern is magnified by the fact that IDV technology can be used as an entry point for other facial recognition systems. All that is needed is for the facial images used in the IDV process to be repurposed to populate a database which is then used in a one-to-many biometric identification system. Such technology is often used in the policing context and has been the subject of considerable public debate—a debate that has been amplified by the use of the technology in Hong Kong to identify and arrest protesters and activists.¹⁴

Now, we are beginning to see governments introduce bans on the use of the technology.¹⁵ Similarly, data protection authorities have begun to recognize the great capacity for harm posed by this technology and are starting to issue fines for its unlawful use.¹⁶ It is therefore critical to continuously examine how the technology is used and whether such use should be permitted.

II. The Technical Requirements to Build an AI Technology to Perform Remote Biometric Identity Verification¹⁷

The machine learning models that power IDV are broadly used for three tasks—(i) to automatically extract data from an identity document;¹⁸ (ii) to validate the identity document and detect digital manipulations of the identity document photo;¹⁹ and (iii) to perform a biometric facial verification between the identity document photo and the live selfie. In performing those tasks, the challenge for the machine learning models is to take an input and produce the correct output at a level equal to or greater than that of a human. However, this is easier said than done.

For the biometric facial verification in particular, the facial recognition technology powering this task has recently experienced a significant boost in accuracy and is capable of performing better than most humans in deciding whether two images show the same person or not. This improvement can be largely explained by three factors—(i) advances in deep learning techniques; (ii) the increased availability of computing power; and (iii) the use of large datasets of facial

¹³ Facial recognition technology can be used to serve three broad use cases—identification, authentication, and verification. Identification is where an individual is identified by comparing their facial image to a database of facial images for the purpose of uniquely identifying the individual. Authentication is where an individual's facial image is matched to a stored identifier (either locally or in the cloud) for the purpose of granting or denying access to a specific system (for example, Apple's FaceID). Verification is used to match two facial images to prove that they belong to the same person, not to ascertain the identity of that person. This nuance in use case is reflected in Article 9 (on the processing of special categories of personal data) of the EU General Data Protection Regulation as only biometric data processed "for the purpose of uniquely identifying a natural person" is considered a special category of personal data. However, the courts and regulators are still working through what it specifically means to uniquely identify a natural person. See the World Bank Group, *Identification for Development (ID4D): Practitioner's Guide* (October 2019), available at <http://documents.worldbank.org/curated/en/248371559325561562/pdf/ID4D-Practitioner-s-Guide.pdf>; See also Commission nationale de l'informatique et des libertés (CNIL): *Reconnaissance Faciale: Pour Un Debat À La Hauteur Des Enjeux*, available at https://www.cnil.fr/sites/default/files/atoms/files/reconnaissance_faciale.pdf.

¹⁴ See *The New York Times*: "In Hong Kong Protests, Faces Become Weapons," (July 26, 2019) available at <https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html>.

¹⁵ See CNN: "Beyond San Francisco, more cities are saying no to facial recognition," (July 17, 2019), available at <https://edition.cnn.com/2019/07/17/tech/cities-ban-facial-recognition/index.html>. The UK has also introduced a bill to temporarily ban the government's use of facial recognition technology until a thorough review can be conducted. See *Automated Facial Recognition Technology (Moratorium and Review) Bill [HL] 2019-20*, available at <https://services.parliament.uk/bills/2019-20/automatedfacialrecognitiontechnologymoratoriumandreview.html>; see also in the United States, *S.2878 - Facial Recognition Technology Warrant Act of 2019*, available at <https://www.congress.gov/bill/116th-congress/senate-bill/2878?r=3&s=1>.

¹⁶ The Swedish data protection authority has fined a school for use of the technology to monitor attendance. See *Datainspektionen*, "Facial recognition in school renders Sweden's first GDPR fine," (August 21, 2019), available at <https://www.datainspektionen.se/nyheter/facial-recognition-in-school-renders-swedens-first-gdpr-fine/>.

¹⁷ This section is co-authored by Martins Bruveris. Martins is a machine learning research scientist in the biometrics team at Onfido Limited. Prior to joining Onfido, Martins was a lecturer in the Department of Mathematics at Brunel University London and a postdoctoral researcher at the École Polytechnique Fédérale de Lausanne.

¹⁸ Automatic data extraction requires the AI technology to recognize the large variety of document types in use worldwide. After determining the identity document type and extracting the data such as name and date of birth, the technology will then compare that information against the information contained in the machine-readable zone of the document to validate the document. If the data printed on the document does not match the information contained in the machine-readable zone, there is likely fraud.

¹⁹ A common fraud technique is to digitally manipulate a photo of an original document by changing the name, date of birth, or swapping the picture.

images. Combined, these three factors enable the training of machine learning models that measure the similarity of two facial images and make an automated decision that is correct in more than 99.9% of cases.²⁰ However, to achieve and maintain such a level of performance in a production system requires a considerable amount of on-going work as we will explore below.

Now let's look at what happens underneath this determination of whether two images are sufficiently similar. At a high level, all facial recognition algorithms work the same way.²¹ Given a dataset of facial images, the machine learning model is trained to map a facial image to a numerical feature vector.²² Images of the same person are mapped to vectors with a high similarity score while images of different people are mapped to vectors with a low similarity score. To determine whether two images show the same person, a similarity score between the corresponding feature vectors is computed and a decision is made based on a designated threshold. If the similarity score is above the threshold, images are considered to be of the same person, but if the score is below the threshold, images are considered to show two different people and the facial similarity check would likely fail. However, if the similarity score is a perfect match or near perfect match, then the exact same photos are likely being compared to one another which is treated as a fraud signal for the purpose of a facial verification where the two images are expected to be different (e.g., the photo within the identity document compared to the live selfie for IDV purposes).

What a model can learn is dependent upon the data that used to train it.

The threshold chosen is based on a trade-off between two kinds of errors the machine learning model can make. A high threshold will require two feature vectors to be very similar to be accepted and may therefore lead to a high *False Rejection Rate (FRR)* where image pairs showing the same person are more likely to be rejected. Conversely, a low threshold may lead to a higher *False Acceptance Rate (FAR)* where image pairs showing two different people are more likely to be accepted. How to resolve this tension will depend on the specific use case and given risk tolerances of such use case, and consequently, biometric facial verification technologies should incorporate configurable thresholds. For example, an IDV used to identify a person taking out a large mortgage may be higher risk than an IDV used to identify a peer-to-peer ridesharing passenger. In those circumstances, the mortgage lender may decide to set a higher threshold than the peer-to-peer ridesharing company.

Regardless of the threshold, what a model can learn is dependent upon the data that is used to train it. In most cases, a model will be more accurate in assessing images that are similar to those that were used during its training.²³ Therefore, in order to ensure that facial verification technology is operating correctly and for all types of individuals, it is necessary to ensure that the training data set is sufficiently large and diverse to be representative of the broad range of images which will be processed during an actual facial verification. While the exact number of images needed to populate such a training set is not concretely known, past studies have demonstrated that a training set in the high millions may be required.²⁴ These demanding requirements push the boundaries on what might be possible using a data trust or other data sharing model as it may be difficult to obtain personal data that satisfactorily reflects the personal data processed in a particular application and in the necessary

²⁰ The top-ranked algorithm in the United States National Institute of Standards and Technology's (NIST) evaluation of facial recognition technology correctly identifies 99.9% pairs of matching passport photos while mistaking only 0.01% of non-matching passport photos as showing the same person. See NIST FRVT 1:1 Verification, available at <https://pages.nist.gov/frvt/html/frvt11.html>.

²¹ For an overview of modern approaches to facial recognition technology, see *Deep Face Recognition: A Survey*, Mei Wang, Weihong Deng, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (February 12, 2019), available at <https://arxiv.org/pdf/1804.06655.pdf>.

²² A numerical feature vector is a list of numbers, usually several hundred long. Each number on its own has no specific meaning, but when taken together, they are sufficient to represent the unique structure of the face. However, it is not possible to use a numerical feature vector to identify an individual without having access to either (i) the algorithm used to generate the numerical feature vector or (ii) a database of feature vectors of already identified individuals computed using the same algorithm.

²³ As an example, machine learning researchers were able to fool an image classification system by inserting familiar objects such as a motorcycle in images in unusual positions, for example by placing the motorcycle upside down. Consequently, the researchers significantly reduced the accuracy of the algorithm. See *Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects*, Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, Anh Nguyen (April 18, 2019), available at <https://arxiv.org/pdf/1811.11553.pdf>. See also *ZDNET: Google's image recognition AI fooled by new tricks* (November 30, 2018), available at <https://www.zdnet.com/article/googles-best-image-recognition-system-flummoxed-by-fakes/>.

²⁴ While causation is very difficult to establish, the steady improvement of results on the Labelled Faces in the Wild (LFW) benchmark, one of standard facial recognition benchmarks, in the years 2010-2015 coincides with increases in the size of training datasets. While size is certainly an important factor, other relevant aspects are the number of distinct identities in the dataset, the variety of images, and the percentage of incorrect labels. See *Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?*, Erjin Zhou, Zhimin Cao, Qi Yin (January 20, 2015), available at <https://arxiv.org/pdf/1501.04690.pdf>. See also *VGGFace2: A dataset for recognising faces across pose and age*, Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi and Andrew Zisserman, Visual Geometry Group, Department of Engineering Science, University of Oxford (May 13, 2018), available at <https://arxiv.org/pdf/1710.08092.pdf>.

volumes.²⁵ Nevertheless, having access to a shared, diverse, and large-scale dataset would provide a useful basis to build facial verification technology using techniques such as transfer learning.²⁶

Diversity in facial images arises from two main sources—extrinsic and intrinsic factors. Extrinsic factors such as camera type, image resolution, illumination, facial expression, and head pose reflect the environment where the photo was captured and are independent of the characteristics of the person. Intrinsic factors such as age, gender, ethnicity, skin tone, and facial geometry are tied to the characteristics of a person and cannot be changed at a given moment in time. There are other factors that sit somewhere between these categories. For example, the security features and printing method of a document significantly impacts the appearance of the identity document photo, but each person has only access to a small number of different identity documents.

If different kinds of extrinsic or intrinsic factors are under-represented in the training set, the trained machine learning model may suffer from poor performance with respect to the under-represented factors. This is because the training data set is fundamentally unbalanced. For instance, if a model is trained using mostly frontal images of faces, one can reasonably expect that the performance of such a model would be poor when used to assess the similarity of faces with a non-frontal (profile) pose. For IDV technology, most facial images are frontal so lower accuracy for non-frontal images may not be a limiting factor.

However, when the lack of diversity is with respect to the intrinsic factors discussed above, the trained model might show differential performance on subpopulations that are representative of those factors; this is often referred to as algorithmic bias. The performance differential can manifest itself in one of two ways. When presented with images of

individuals from the given subpopulation, a model can be more likely to reject two images showing the same person, resulting in a higher FRR for this subpopulation. Alternatively, the model may more likely accept a pair of images showing different individuals as a match, resulting in a higher FAR. The consequences in each case are different.²⁷

- A higher FRR for a subpopulation would mean that more individuals are rejected by the automatic process and face disparate treatment, and the companies using the IDV technology will onboard less individuals.
- A higher FAR, in contrast, would mean that individuals are at greater risk of having their identity stolen, and the companies using the IDV technology will have greater exposure to fraud.²⁸

In other words, a machine learning model is often described as a tunable machine that requires continuous monitoring and correction to ensure it is performing with high levels of accuracy. If the machine learning models are not working properly, there can be significant consequences for individuals. This is particularly true for IDV as an incorrect output can mean that a person is denied access to vital services that they should be rightfully able to access or could have their identity stolen.

A machine learning model is often described as a tunable machine that requires continuous monitoring and correction to ensure it is performing with high levels of accuracy.

²⁵ For an overview of data trusts, see the Open Data Institute's Data trusts: lessons from three pilots (April 15, 2019), available at <https://theodi.org/article/odi-data-trusts-report/>.

²⁶ For example, to develop a facial recognition algorithm that could compare selfies with passport photos, researchers first trained a facial recognition algorithm on a dataset of "wild" images followed by training on a task-specific dataset of selfie/passport pairs. The resulting facial recognition algorithm achieved greater accuracy than would have been possible by training on either of the datasets separately. See Large-scale Bisample Learning on ID Versus Spot Face Recognition, Xiangyu Zhu, Hao Liu, Zhen Lei, Hailin Shi, Fan Yang, Dong Yi, Guojun Qi, Stan Z. Li (February 14, 2019), available at <https://arxiv.org/pdf/1806.03018.pdf>.

²⁷ For example, machine learning researchers have observed that different facial recognition technologies have different error rates for different ethnicities. It is likely that data imbalance is one cause of this differential performance, but further research is needed to establish whether it is the only cause. See Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network, Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, Yaohai Huang, Beijing University of Posts and Telecommunications, Canon Information Technology (Beijing) Co., Ltd (July 27, 2019), available at <https://arxiv.org/pdf/1812.00194.pdf>.

²⁸ Importantly, the implications of a higher false rejection rate and a higher false acceptance rate described just above are unique to identity verification where an individual is seeking access to a service and the identity verification is used to prove ownership of their presented identity document. Where facial recognition technology is used differently, for example, to identify a single individual out of the public, the implications for the individual are reversed as a higher false acceptance rate means that the individual is then likely to be incorrectly identified and subject to disparate treatment. For more information about the difference between a one-to-many identification as compared to a one-to-one verification, see footnote 14 above.

III. The New Privacy Challenges of Using Personal Data to Build AI Technologies

If, as we have established, it is true that personal data has become the bricks and mortar used to build AI technologies, then we are presented with an ethical dilemma. On one hand, the data protection and privacy rights of the individual whose personal data is used to build the AI technology must be protected and safeguarded. On the other hand, the AI technology will only be useful if it is trained to perform at the necessary high levels of accuracy to ensure that the individuals using the AI technology are not unfairly prejudiced or otherwise harmed. Enshrined in both objectives is the need to preserve and protect the fundamental human rights of the individuals involved in the building and use of AI technologies.²⁹

As mentioned previously, the purpose for processing personal data to build AI technologies is not to affect the individual whose personal data is being processed, but rather, to build the AI technology in such a way where it will perform with a high degree of accuracy. While such processing is necessary to ensure that the AI technology operates correctly, the individual whose personal data is used to build the AI technology does not necessarily benefit directly from the technology's enhanced capabilities. There is an asymmetry. The individual's personal data is retained, and the machine learning models are trained on it over time. Yet, the direct benefit to the individual is perceivably low.

The entities building AI technologies are also frequently not the same entities that are deploying the AI technologies, and in these circumstances, the company building the AI technology may have no direct relationship with individuals. This is seen with IDV where the majority of entities using the AI technology are capturing the images of the individual's identity document and face and then sending those images to the IDV provider directly via an API call.³⁰ And where the IDV provider does provide the image capture experience to the individual, the individual's direct interaction with the IDV provider is often limited to the image capture screens. Even this perceived interaction is also frequently removed because the service has been white labelled.³¹ What this means is that the IDV provider

has very limited opportunities to meaningfully interact with the individual whose personal data is being processed.

The inability for companies building AI technologies to interact with individuals is common, and these companies will frequently seek to meet their data privacy legal obligations by passing those obligations onto the companies deploying the AI technology via contract. Such contractual clauses will typically oblige the companies deploying the AI technology to provide individuals with all necessary notices, obtain all necessary consents, and otherwise do all that is required for the AI technology provider to provide and build the AI technology.

If it is true that personal data has become the bricks and mortar used to build AI technologies, then we are presented with an ethical dilemma.

While the company building the AI technology has contractually passed on their legal obligations to the company deploying the AI technology, the reality is that such company may not be fulfilling those passed on obligations. As a result, the company that built the AI technology may have a "defensible position" under the law, but that company is likely not fulfilling the spirit of the law and meeting their ethical obligations. The unfortunate outcome is that the individuals using the AI technology are not properly notified of how their personal data is being used. Nor are those individuals empowered to control that data usage.

This shift in how personal data is used to build AI technologies, combined with the complexity over the personal data supply chain inherent to many AI technologies, has created new privacy challenges. At a fundamental level, it is necessary to re-examine the data protection and privacy principles that are used as the backbone to many of the data protection legal frameworks around the world.³² These principles include:

²⁹ As an example, compare Article 7 (Non-Discrimination) to Article 14 (Right to Privacy) of the United Nations Universal Declaration of Human Rights, available at <https://www.un.org/en/universal-declaration-human-rights/>. Similarly, compare Article 14 (Prohibition of Discrimination) to Article 8 (Right to respect for private and family life) of the European Convention on Human Rights, available at https://www.echr.coe.int/Documents/Convention_ENG.pdf.

³⁰ An "API" is an "application programming interface" and is a means to request and exchange data between data services.

³¹ To white label a service is where a company re-brands a service as their own to give the perception that they have built and are operating the service. This is a common activity designed to provide a uniform user experience but does limit overall transparency.

³² See Article 5 of the EU GDPR (footnote 4). See also the OECD Privacy Guidelines (footnote 5). See also the PIPEDA Fair Information Principles, available at https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/.

The complexity of the AI supply chain has created an ecosystem where legal compliance may be demonstrated but the spirit of the law is often not fulfilled.

- *Data Minimization* - Personal data must be adequate, relevant, and limited to what is necessary to fulfill the purpose for processing. However, when that purpose is to build AI technologies, it is unclear when that purpose is fulfilled. How much personal data is required for a model to perform at the necessary levels of accuracy? When the intention is to build an AI technology, it is often difficult to know in advance how much personal data will be necessary, which can lead to the temptation to capture as much personal data as possible. Similarly, advancements in AI technology may enable the same goal to be accomplished using less personal data or anonymized data.
- *Retention Limitation* - Personal data must not be stored and processed for a period of time longer than is necessary to fulfill the purpose for processing. However, as applied to processing for the purpose of building AI technologies, that purpose is frequently on-going and without a clear end date. In contrast to when personal data is processed to provide a requested service with a clear finite end date, personal data continues to have utility when being processed to build an AI technology so long as that personal data is reflective of the inputs processed by that AI technology when deployed in the real world. For example, facial images will continue to be useful for training facial verification models so long as people continue to look like people.³³
- *Transparency* - Individuals must be notified of how their personal data will be processed. This is not only to inform individuals but also to empower individuals to make decisions about the use of their personal data. However, when using personal data to build AI technologies, the use of an individual's personal data is not intended to impact them at all.³⁴ This can be a very difficult idea to communicate and the logical conclusion for most people would be to not have their personal data processed if there is no intended impact on them. More should be done to educate the general public on how AI technologies are built.
- *Control and Choice* - Individuals have the fundamental right to control how their personal data is used. This is often thought of in the context of opt-in consent, but where many millions of data samples are required to build an AI technology, is opt-in consent a viable solution? Do alternative mechanisms exist that provide individual control and choice while still enabling the processing of the large volumes of data necessary to build AI technologies?
- *Data Quality* - In order to ensure that an individual is treated fairly and correctly, it is necessary to only process accurate data and to take every reasonable step to correct inaccurate data. Yet, when building AI technologies, inaccurate or poor-quality data will not result in a direct negative outcome for the individual whose personal data is processed to build the technology. Instead, the accuracy and performance of the overall AI technology will be degraded and every person who is then subject to the AI technology will suffer. It is therefore important that the most relevant and accurate data be used to train a machine learning model.
- *Accountability* - Organizations processing personal data must be able to prove and demonstrate that they have met their compliance obligations. Yet, the complexity of the AI supply chain has created an ecosystem where legal compliance may be demonstrated but the spirit of the law is often not fulfilled. More should be done to prove and demonstrate ethical compliance to ensure that the fundamental human rights of individuals are protected and safeguarded.

³³ Research has shown that artificial intelligence machine learning models can exhibit a behavior termed catastrophic forgetting where a model trained first on dataset A and then on dataset B will perform worse on dataset A at the end of the training compared to a model that was trained only on dataset A. The conclusion being that if performance on dataset A remains important, then the model should be trained on both A and B, which of course requires continued access to dataset A. See Measuring Catastrophic Forgetting in Neural Networks, Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, Christopher Kanan, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology (November 9, 2017), available at <https://arxiv.org/pdf/1708.02072.pdf>.

³⁴ Yet, once the technology is created, there is the potential for the technology to directly impact the relevant person if the technology is used towards that person.

IV. Ethical Solutions to Using Personal Data to Build AI Technologies

As technology continues to evolve, not only must the law keep pace, but how the law is applied must also keep pace. To do so, ethical solutions are needed that reflect the spirit of law. When it comes to privacy, this means finding solutions that are in the interests and benefit of the humans that interact with the technology—including not only those individuals who use the AI technology but also those individuals whose personal data is used to build the AI technology.

While existing legal and ethical frameworks provide clear boundaries and directions as to how personal data may be processed, those boundaries and directions are less clear when it comes to using personal data to build AI technologies. Further guidance is needed.

To better help organizations understand how to meet the requirements of those legal and ethical frameworks when processing personal data to build AI technologies, the following principles are meant to provide a path forward:

1. *Acceptable Data Use* - If opt-in consent is not viable, then personal data must only be used to build AI technologies where the AI technology satisfies a legitimate societal interest or public benefit.
2. *Individual Empowerment* - Individuals must be empowered to understand and control how their personal data is being used at different points in the AI supply chain.
3. *Data Architecture* - AI technologies must be built securely and in such a manner to facilitate individual empowerment.

However, it is important to recognize that even if organizations adhere to these principles when building AI technologies, the issue of the legal and ethical use of these technologies is not addressed. How AI technologies are deployed in individual

use cases is extremely important, and as discussed above, there are a number of existing and developing ethical frameworks that address this issue. To touch on each of the above principles in more detail:

Acceptable Data Use

As demonstrated above with IDV, the building of many AI technologies—particularly those using machine learning or deep neural networks—depends on processing large and diverse data sets that include data samples in the high millions. This data requirement makes it exceptionally difficult, if not impossible, to obtain the necessary data volumes on an opt-in consent basis where the data samples must include personal data.

Much of the difficulty in obtaining opt-in consent can be traced back to what has been described as status quo bias.³⁵ When presented with a choice over how personal data is used, studies have shown that individuals will frequently keep the default setting. This is why it is so important that privacy by design requires privacy-preserving default configurations. Yet, such default configurations may also mean that it is not possible to collect the personal data volumes necessary to build many AI technologies on an opt-in consent basis as individuals will frequently not take the time to understand or change the default setting to permit the personal data usage, even if the individual agrees with that personal data usage. The implications of this are a lack of personal data to properly build AI technologies.³⁶

If opt-in consent is not viable, then the data must not be used unless an alternative lawful basis is identified. This concept of lawful basis exists in different forms in different data protection and privacy laws around the world, but at its most basic, it is the search for whether or not the data usage is an activity that society deems acceptable and commonly requires a balancing of interests.³⁷

³⁵ Acquisti, Alessandro and Adjerid, Idris and Balebako, Rebecca Hunt and Brandimarte, Laura and Cranor, Lorrie Faith and Komanduri, Saranga and Leon, Pedro and Sadeh, Norman and Schaub, Florian and Sleeper, Manya and Wang, Yang and Wilson, Shomir, Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online (August 7, 2017). A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. Cranor, S. Komanduri, P. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. ACM Comput. Surv. 50, 3, Article 44 (August 2017). Available at SSRN: <https://ssrn.com/abstract=2859227> or <http://dx.doi.org/10.2139/ssrn.2859227>.

³⁶ Similar phenomena have been demonstrated in the context of organ donation whereby countries with a default organ sharing program have far higher rates of organ donation. See Eric J. Johnson and Daniel G. Goldstein, "Defaults and Donation Decisions," *Transplantation* 78 no. 12 (2004), available at <https://www.ncbi.nlm.nih.gov/pubmed/15614141>.

³⁷ As an example, see Article 6 of the EU GDPR (footnote 4) which provides several alternatives to lawfully process personal data without an explicit opt-in consent. These include where the processing is in the legitimate interest of the controller or other third party as well as where the processing is in the public interest. For an example of a balancing test, see the UK ICO's Legitimate Interest Assessment template, available at <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/>.

However, not all personal data is equal. Some personal data is considered riskier than others and requires heightened protections and safeguards. As an example, racial and ethnic data is considered high risk in many legislative frameworks.³⁸ This is of course logical in that racial and ethnic data should only be used to inform a decision about a person when absolutely necessary or with the individual's consent. This is to avoid unfair and prejudicial treatment.

Yet, some of the most dangerous risks in AI technologies, such as algorithmic bias, cannot be sufficiently addressed without the processing of racial and ethnic data. While much of the required machine learning research will not necessitate identifying individuals on the basis of race and ethnicity, both a test data set and a validation data set containing racial and ethnic labels does need to be created to measure the algorithmic bias and then, to measure the mitigation of such bias.

In deciding whether the processing of personal data to build an AI technology should be permissible absent an opt-in consent, the following questions should be asked and the answers balanced against one another:

- Is anonymization possible?
- Is the AI technology being built using machine learning, deep neural networks, or another subset of AI which requires large and diverse data sets such that consent is not viable?
- Is the personal data high risk?
- Are individuals likely to object to the use of their personal data?
- What harms might individuals suffer if the personal data is lost, stolen, or otherwise used for an unintended purpose?
- Who will benefit from the creation of the AI technology?
- Will the AI technology serve a societal interest or public benefit?
- What harms might the public suffer if the AI technology is not available?

- What harms might the public suffer if the AI technology is available but performing at a degraded level?
- What harms might the public suffer if the AI technology is used for an unlawful purpose or otherwise abused?

The above questions and balancing test are merely intended to supplement existing data privacy legal and ethical frameworks, and to the extent that such frameworks explicitly say how personal data should be used to build a given AI technology, that framework should control.

Individual Empowerment

If personal data is going to be used to build an AI technology without first collecting opt-in consent, everything must be done to further empower individuals to understand and control how their personal data is used. Yet this is far easier said than done, and the complexity is in the supply chain of data.

As discussed above, the company building the AI technology is frequently not the same company that is making the AI technology available. This reality requires the company building the AI technology to provide the companies using the AI technology with clear information about how the AI technology functions, performs, and can be controlled.³⁹ Though, even having provided the companies that are using the AI technology with the necessary information, there is no guarantee that those companies will empower individuals with the needed notice and controls.

To ensure that the necessary notices and controls are reaching the impacted individuals, the company building the AI technology could review the user flows, privacy notices, and controls of those companies using the AI technology. However, the workload associated with reviewing hundreds or thousands of user flows, privacy notices, and controls as well as the inherent risk of providing another company with legal advice makes this task likely unachievable for most companies. In essence, the company building the AI technology would be taking on the role of a privatized regulator. As a mitigation, the company building the AI technology should, at a minimum, undertake a general due diligence / know your business customer exercise to understand whether the company intending to use the AI technology is reputable and likely to fulfill their legal obligations, including the

³⁸ See Article 9 of the EU GDPR (footnote 4).

³⁹ This information should include not only privacy notices but also product specifications, privacy impact assessments, and information on the statistical performance of the AI technology. In generating this information, a helpful exercise is to perform user testing with both the company using the AI technology and the individuals whose personal data will be processed by the AI technology. Absent providing this education through these materials, there is little hope of the individual receiving fair notice of processing. In addition, many organizations are now also looking at different methods to provide privacy related information, such as infographics, videos, and similar. This is often part of a larger initiative, increasingly known as "Legal Design". See Legal Geek: "Legal Design WTF?" (March 22, 2018), available at <https://www.legalgeek.co/learn/legal-design-wtf/>.

forementioned privacy obligations.⁴⁰ If the prospective customer is not able to pass such a review, then the company building the AI technology should not make the sale.

To further empower individuals, companies building AI technologies should seek to provide as many touch points with individuals as possible. Not only should individuals be able to obtain fair notice and control from the company deploying the AI technology, but the company building the AI technology should also prepare its own privacy notice and make controls available directly to individuals.⁴¹ And of course, any signal to not use personal data for the development of the AI technology should be honored against all data sources—meaning that once an individual expresses their wish to not have their personal data used to build an AI technology, if that individual's personal data is seen again at a later time, that initial opt-out should be honored against that new data source.⁴²

As an additional precaution, companies should consider implementing a delay period to using personal data to build AI technologies—noting that the data processing associated with the provision of an AI technology is frequently entirely separate from the data processing that is used to build the AI technology. This delay in data usage can provide an individual with a period of time in which they can express their wish to not have their personal data used for the purpose of building the AI technology prior to their data ever being used. In effect, this is a compromise solution to obtaining opt-in consent, enabling the use of personal data to build AI technologies while still providing privacy choice and control to the individual.⁴³

Data Architecture

When personal data is used to build AI technologies, there is no intended effect on the individual as to that specific data usage, and as such, the direct risk to the individual is that their personal data is lost, stolen or otherwise used for an unintended purpose to the detriment of the individual. It is

therefore necessary that the computing architecture used to build AI technologies is built in such a way as to ensure that data use choices are honored and personal data is protected with the utmost security.

Such computing architectures must satisfy the following three conditions:

1. *Data Traceability* - Personal data must remain fully traceable back to the individual to whom the personal data relates. This is necessary to ensure that any requests to delete or otherwise control that data can be honored. As discussed above, such requests might come from the companies deploying the AI technology or from the individuals themselves.
2. *Security* - Security is an absolute necessity as a failing in security creates the conditions where an individual is most likely to suffer harm. No researcher should have the ability to freely export personal data from a research environment or use the personal data for an unauthorized purpose. While there are many approaches to security, in these circumstances, the goal is to create a protected and monitored environment with exceptionally tight controls over the import and export of personal data as well as access to the environment itself.
3. *Computing Environment* - While the underlying architecture must support data traceability and have high levels of security, if the computing environment provided to researchers does not meet their needs, they will not use it and will seek out other solutions. Consequently, this shortcoming alone creates a security risk, which is wholly unacceptable. To meet the researchers needs, the computing environment must have sufficient computing power and ease of use. In other words, researchers should be able to conduct research as quickly and easily as they would expect in other, less controlled environments (for example, on local machines).

⁴⁰ A know your business (KYB) privacy review might include checking whether the company has an up to date privacy policy and whether they have been the subject of any privacy enforcement actions or civil litigation, among other standard KYB checks.

⁴¹ As the company building the AI technology likely does not have a direct relationship with the individual, there needs to be a way for the individual to prove their identity before honoring their request. If the individual is requesting a copy of their personal data, then an identity verification, as described within, is likely necessary to ensure that personal data is only being given to the correct person. Yet, if the request is to have personal data removed from an AI data set, then a less privacy invasive identification system should be used as the risk of deleting too much data does not likely justify the use of the identity verification. This concept of only carrying out an identity document-based identity verification request where reasonably necessary has been recognized in a decision by the Danish data protection authority. See Datatilsynet, "ID-validering ifm. anmodninger om udøvelse af registreredes rettigheder," (October 25, 2019), available at <https://www.datatilsynet.dk/tilsyn-og-afgoerelser/afgoerelser/2019/okt/id-validering-ifm-anmodninger-om-udoevelse-af-registreredes-rettigheder/>.

⁴² The irony is that a minimal amount of personal data will need to be retained to maintain the opt-out list.

⁴³ Though, where a company has an automated model retraining pipeline in place that will quickly use the data to retrain the machine learning models on a high frequency basis (possibly daily), then it is likely more in the interests of the individual to quickly use the personal data that one time and no more. Yet, this approach is dependent upon the personal data not being needed to retrain the machine learning model in the future, which is frequently not the case.

AI technologies have the capacity to do a lot of good in the world, but whether they do so is highly dependent upon both how we use and build those AI technologies in the first place.

By achieving the three broad objectives above, machine learning researchers should be able to use an environment that enables machine learning research while protecting the privacy and security of the individual. The ask is not that machine learning researchers become privacy experts but that they are provided a safe space to do what they do best—research.

V. Conclusion

AI technologies have the capacity to do a lot of good in the world, but whether they do so is highly dependent upon both how we use and build those AI technologies in the first place. The unfortunate truth is that personal data has become the bricks and mortar used to build many of these AI technologies and more thought needs to be given as to how this personal data is properly protected and safeguarded.

When faced with this truth, society has the option to (i) decide that such AI technologies should not be built; or (ii)

seek out privacy preserving methods to build these AI technologies. What is not an option is to ignore this truth and expect the AI technology to self-manifest without the required engineering and research development work along with their related privacy impacts on individuals. Both options must be thoughtfully addressed by governments, regulators, the companies using and building these technologies, and the wider public.

To use personal data to build AI technologies requires a re-examination of existing legal and ethical principles and frameworks as applied to the facts of the given research project. Only through a solid understanding of the AI technology that is being built can the privacy choices be well understood. At such time, efforts must then be made to ensure that all is done to protect the individuals whose personal data is used to build the AI technologies. This is achieved through assessing the value of the AI technology, empowering individuals, and by establishing a secure and robust data architecture in which to build the AI technology.



Carr Center Discussion Paper Series

**Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138**

Statements and views expressed in this report are solely those of the author and do not imply endorsement by Harvard University, the Harvard Kennedy School, or the Carr Center for Human Rights Policy.

Copyright 2020, President and Fellows of Harvard College
Printed in the United States of America

**This publication was published by the Carr Center for
Human Rights Policy at the John F. Kennedy School of
Government at Harvard University**

Copyright 2020, President and Fellows of Harvard College
Printed in the United States of America

79 JFK Street
Cambridge, MA 02138

617.495.5819
<https://carrcenter.hks.harvard.edu>

