As artificial intelligence applications become more widespread, it’s of paramount importance that we ensure that privacy and security concerns are well thought through. Companies, governments, policies and all other areas of businesses and research departments must be aware of concerns and regulations. Throughout this paper, we’re taking a look at some of the key concerns with the goal of understanding the risks involved, and how we can ensure AI is used safely and with both personal and corporate privacy at the helm.

The paper features contributions from experts working in the field, featuring their research and real world applications. We explore not only the potential threats and dangers, but also the ways in which AI can be leveraged to improve personal privacy and security systems through smarter models, improved regulations, policies and more. Learn from global experts from the likes of Microsoft, McGill University, Google Brain and more.

CONTENTS

Contents..........................2 - 3
Introduction .....................4 - 5
Current Threats...............6 - 9
Frameworks & Regulations...10 - 12
AI Enhancing Privacy........13 - 14
Case Study: Identity Fraud...15 - 16
Case Study: Government....17 - 18
Preparing for the Future.....19 - 20
Learn More.......................21
1. CONTRIBUTORS

**Mariana Acuna**, Co-Founder, Opaque Studios
Key focuses are on AI, VR, MR and AR - Opaque Studios are shaping the future of filmmaking by building tools that help realize the Director's vision. Mariana studied ‘Artificial Intelligence: Implications for Business Strategy’ at Massachusetts Institute of Technology.

**Alison Darcy**, CEO & Founder, Woebot
A clinical research psychologist, Alison was Faculty in Psychiatry and Behavioural Sciences at the Stanford School of Medicine for a decade where she developed digital health interventions for young people. With both a technical and scientific background, Alison works at the intersection of mental health and technology.

**Eli David**, CTO & Co-Founder, Deep Instinct
Specialist in deep learning and evolutionary computation with over thirty publications in leading AI journals. Eli is a developer of Falcon, a chess playing program based on genetic algorithms and Deep Learning. He received the Best Paper Award in 2008 Genetic & Evolutionary Computation Conference.

**Ian Goodfellow**, Research Scientist, Google Brain
Ian's research interests include most deep learning topics, especially generative models and machine learning security and privacy. He invented generative adversarial networks, was an influential early researcher studying adversarial examples, and is the lead author of the MIT Press textbook Deep Learning (www.deeplearningbook.org). He runs the Self-Organizing Conference on Machine Learning, which was founded at OpenAI in 2016.

**Abhishek Gupta**, AI Ethics Researcher & Software Engineer, Microsoft, Montreal AI Ethics Institute and McGill University
Prestige Scholar from McGill University with a degree in Computer Science and Economics, Abhishek is an AI Ethics Researcher at McGill University and the lead organizer for the Montreal AI Ethics meetup. Most recently, Abhishek has written a paper on how large-scale data breaches combined with advances in deep learning techniques will lead to the next generation of identity theft and fraud.

**Yuanyuan Liu**, Statistical Machine Learning, AIG
Yuanyuan is working on AIG’s innovative R&D projects to apply deep learning algorithms in insurance and investment, using generative models, sequential models, computer vision, and reinforcement learning. Yuanyuan graduated from Oxford with a DPhil in Statistical Data Mining and an MSc in Applied Statistics.

**Jeremy Marvel**, Research Scientist, U.S. National Institute of Standards and Technology (NIST)
Research expertise include intelligent and adaptive solutions for robot applications, with particular attention paid to human-robot and robot-robot collaborations, multirobot coordination, industrial robot safety, machine learning, perception, and automated parameter optimization.
Fiona McEvoy, Founder, YouTheData
With a graduate degree in Philosophy, Fiona’s key research is in ethics and technology. She recently presented her work to the International Association for Computing and Philosophy (IACAP) at Stanford University, and her latest paper on Big Data and the future of autonomy was published by the University of Bucharest as part of a series on emerging technology.

Andrew McStay, Professor of Digital Life, Bangor University
Author of a range of books on advertising, creativity, privacy and the social impact of new technologies. Supported by the Arts and Humanities Research Council, he has conducted extensive citizen-based fieldwork and interviews to assess the potential impact of technologies that profile emotional life. He’s currently writing a book ‘Emotional AI: The Rise of Empathic Media’.

Anton Prokopyev, Data Scientist, The World Bank
Specializes in Natural Language Processing and Text Analytics using R. Having received his master’s degree from University of California, Anton went on to work at some of the Bay Area’s most prominent technology and online startup companies.

Andrea Renda, Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy, Centre for European Policy Studies
Research lies at the crossroads between economics, law, technology and public policy. Andrea is Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy at the Centre for European Policy Studies (CEPS).

Toby Simpson, CTO & Co-Founder, Fetch.AI
Toby has over a quarter of a century of experience in software architecture and development and more than a decade as CTO across three companies. As a software developer and manager, he’s built and directed several successful computer games including producing the highly successful Creatures series in the 90s.

Clint Wheelock, Founder & Managing Director, Tractica
Extensive background in market intelligence focused on emerging technologies. Most recently, he was founder and president of Pike Research, a leading market intelligence firm focused on the global clean technology industry, which was acquired by Navigant Consulting.
2. INTRODUCTION

Andrew McStay, Professor of Digital Life, Bangor University, 2018

Privacy and security. Two buzzwords that we hear a lot of, but why all the concern, and what do we actually mean by these frequently and sometimes freely used terms?

Privacy

The recurring suggestion that privacy might be dropped from the human equation is to misunderstand it. This is because it is best understood by turning away from screens to recognise that privacy plays a fundamental role in daily interactions, what is considered taboo, intimacy, the confidences we share with others, how we arrange domestic and other spaces, and where we store thoughts and things of value.

Privacy is not a thing but an ethical protocol that governs interaction, relationships and behaviour in given situations.

Beyond its centrality in governing human interaction, and relationships with machines and organizations, the ethical components of privacy are also important. From liberal politics and philosophy around the 17th and 18th centuries, it is constructed out of notions such as freedom, consent, autonomy, self-determination, dignity and non-interference from unwanted others.

Privacy is personally, collectively, politically and ethically important.

Security

Isn’t privacy another word for security?

In the domain of data protection, privacy can be seen as a parental principle to security. While privacy certainly has to do with keeping personal and sensitive information safe and secure (be these hand-written letters, encrypted financial information or a nation’s health data), the ethical building blocks of privacy causes it to be more than prevention of access by unwanted outsiders.

One way to think about the relationship is that security is about means, while privacy is about ends. For example, encryption is a security practice, but privacy is the right and freedom for people to communicate without fear of interception. Admittedly, privacy and ethics-by-design blurs the privacy/security and end/means distinction, but the argument that privacy is parental to security is sound.

As artificial intelligence (AI) and machine learning (ML) are increasingly applied to intimate dimensions of human life, they invoke a clear need to respect dignity and see people as ends rather than means. If this guiding principle were borne in mind and practice, there is scope to have the best of AI/ML and less of the dystopian.

Throughout this paper, we will refer to the terms security and privacy as follows:

SECURITY: Protecting individuals or organisations from using your personal information against you, e.g. health records, personal finances, social security details

PRIVACY: Individuals secluding themselves from the public eye.
Is privacy about hiding things away?
Sometimes yes, but mostly no. Privacy should not be equated with seclusion and hiding. People are willing to share the most intimate of details about themselves under the right conditions. This means it isn’t a paradox to say that one can be highly open about one’s life yet enjoy privacy. It’s about respecting norms, conventions and choices, but perhaps foremost, dignity.

“This means privacy is dynamic in that what people will share is subject to regional, circumstantial and historical change, but what doesn’t change is respect for human rights, dignity and agreed principles of data sharing with others (be these friends or organizations).

How is AI and ML altering privacy?
AI will continue to shape human life in profound ways. To give an example, Amazon’s Rekognition grants capability for all sorts of organizations to recognize and analyze objects, people, text, scenes and activities, in images and video. Computer vision has scope for social good, but also invokes seismic privacy questions. Foremost in the case of Amazon’s service is societal questions about the desirability of automated mass visual surveillance and loss of public anonymity. The significance of this is an undemocratic transference of power from citizens to governments. While debate regarding balance between liberty and security is old, AI raises concerns about an increase of power over society.

Privacy, for who?
AI applications are increasingly at the heart of modern institutions. As hidden algorithms increasingly shape important parts of our lives (loans, health predictions, employment chances, prison sentencing), it has become hard for citizens to challenge faulty decisions. Of course, human decision-makers consciously and unconsciously make bad decisions, but they are more easily challenged. A key unenviable challenge for services using AI for decision-making is to explain the terms of a decision and provide scope for meaningful redress. Given that AI will make determinations about life chances and opportunities, this transparency is important.

CASE STUDY
Currently, a vast majority of ML services need to push your data to the cloud to perform analysis and put AI into effect. Big players in this space such as Google and Apple are working on chips to carry out these processes directly on the device, resulting in a decreased risk of users data getting leaked or hacked. (James Vincent, The Verge, 2018)

Tractica expects that 2019 and 2020 will be the years when a ramp-up in deep learning chipset volumes will take place and winners will begin to emerge. (Clint Wheelock, Tractica, 2018)
2. CURRENT THREATS

Contrary to what people think, scientific progress is slow and continuous, but social and economic impact can be disruptive - there’s no doubt that we will reach human level AI but we don’t know how long it’s going to take. We need to make sure that AI and society will bloom for the benefit of all.

(Yoshua Bengio, 2017)

As Bengio stated, we need not worry about the distant future of AI; The bigger immediate concern is the misuse of AI through negligence or malice rather than the rise of super-intelligent machines with AGI, as this still remains science fiction.

Engineers are creating increasingly complex machine learning models that are capable of making decisions and taking actions without the assistance of their creator. Whilst the potential for staggering real world applications in multiple industries is evident, much like any piece of software, machine learning models are prone to breaches.

In the paper, Stealing Machine Learning Models via Prediction APIs, computer scientists from Cornell Tech, EPFL and University of North Carolina reverse engineered a machine learning trained AI, based only on sending them queries and analyzing the responses. Whilst ML models may be deemed confidential due to their sensitive training data, commercial value or use in security applications, models are still deployed with publicly accessible query interfaces. This means that in some instances, users are able to train models on potentially sensitive data.

Where are the AI vulnerabilities?

Technology that surrounds us is perpetually recording data and compiling it on a scale never seen before in human history. This data is analyzed and repurposed in a variety of different ways, but problematically it can reveal more about us than we actually intended to reveal. Currently, there is still a dramatic imbalance between what “data gatherers” know, and what “data emitters” (i.e. us) understand of this. This brings with it privacy problems, and current privacy disclaimers are often thought to be an inadequate solution.

Many of us aimlessly click through disclaimers without reading them or understanding the consequences. In truth, we - the public, governments, the tech industry - need to understand what mechanisms can be put in place to protect users and ensure that we don’t blindly assent to the ongoing ransack and scrutiny of everything we do online.

Aside from knowing more about us than we might anticipate, public privacy (and therefore safety) is also highly contingent on the measures data gatherers take to protect our information.

(Fiona McEvoy, YouTheData, 2018)

As researchers have proved that even though they may be creating black box algorithms that are seemingly ‘secret’. the possibility for reverse engineering means that the algorithms can be replicated, hacked or even stolen.
WHERE DO WE GIVE OUT INFORMATION?

- **Home assistants**
  - Voice data
  - Online browsing & purchases
  - FAQs
  - Purchases

- **Location services**
  - Frequently visited locations
  - Home address
  - Journeys

- **Web browsing**
  - Search history
  - Purchases

- **Social media**
  - Photos
  - Conversations
  - Personal details e.g. name, age, education, employment

PERSONAL PRIVACY

Machine and Deep Learning systems require large amounts of data. Some of that data is collected in private spaces like our homes, and users reveal sensitive information to these systems. Think about when you sign up for a social media platform or buy something online - we upload information such as what we like, where we work, where we live amongst many other preferences which in turn allows systems to make suggestions for us. When we interact with these systems and they learn more and more about us, the potential repercussions of breaches becomes more severe.

*Doctors and lawyers are prohibited from using clients’ information for their own interests, so why aren’t Google and Facebook?* (Jack M. Balkin, 2016)

People are increasingly aware that we pay for the benefits of technology with our personal data. Facebook, as we’ve learned in recent months, is not “free” in the way we initially imagined. *In many cases there is a complex exchange occurring between consumers and companies that, in the best of these cases, are mutually agreeable.* Like a tailor taking our (private) measurements in order to make a new suit, we agree to an intrusion so that we may receive something want, like, or need. The problem occurs when private or personal information that we would rather not share is being harvested by companies without our full knowledge.

Now that we live in a world where technology is constantly watching and listening, it needs to be absolutely clear when it is simultaneously collecting for less-than-obvious purposes. There also needs to be more transparency and education with regards to the ways in which our personal information is used to “nudge” us, i.e. when tech companies with a superior knowledge of what we tend to like or want use this understanding in a way that takes advantage of our cognitive flaws and blind spots. Commercial entities have their own objectives, and these don’t always have the protection and privacy of the public as their focus.

*(Fiona McEvoy, YouTheData, 2018)*

AREAS FOR CONCERN

1. Doctors and lawyers are prohibited from using clients’ information for their own interests, so why aren’t Google and Facebook? (Jack M. Balkin, 2016)

People are increasingly aware that we pay for the benefits of technology with our personal data. Facebook, as we’ve learned in recent months, is not “free” in the way we initially imagined. *In many cases there is a complex exchange occurring between consumers and companies that, in the best of these cases, are mutually agreeable.* Like a tailor taking our (private) measurements in order to make a new suit, we agree to an intrusion so that we may receive something want, like, or need. The problem occurs when private or personal information that we would rather not share is being harvested by companies without our full knowledge.

Now that we live in a world where technology is constantly watching and listening, it needs to be absolutely clear when it is simultaneously collecting for less-than-obvious purposes. There also needs to be more transparency and education with regards to the ways in which our personal information is used to “nudge” us, i.e. when tech companies with a superior knowledge of what we tend to like or want use this understanding in a way that takes advantage of our cognitive flaws and blind spots. Commercial entities have their own objectives, and these don’t always have the protection and privacy of the public as their focus.

*(Fiona McEvoy, YouTheData, 2018)*

AREAS FOR CONCERN

1. Doctors and lawyers are prohibited from using clients’ information for their own interests, so why aren’t Google and Facebook? (Jack M. Balkin, 2016)

People are increasingly aware that we pay for the benefits of technology with our personal data. Facebook, as we’ve learned in recent months, is not “free” in the way we initially imagined. *In many cases there is a complex exchange occurring between consumers and companies that, in the best of these cases, are mutually agreeable.* Like a tailor taking our (private) measurements in order to make a new suit, we agree to an intrusion so that we may receive something want, like, or need. The problem occurs when private or personal information that we would rather not share is being harvested by companies without our full knowledge.

Now that we live in a world where technology is constantly watching and listening, it needs to be absolutely clear when it is simultaneously collecting for less-than-obvious purposes. There also needs to be more transparency and education with regards to the ways in which our personal information is used to “nudge” us, i.e. when tech companies with a superior knowledge of what we tend to like or want use this understanding in a way that takes advantage of our cognitive flaws and blind spots. Commercial entities have their own objectives, and these don’t always have the protection and privacy of the public as their focus.

*(Fiona McEvoy, YouTheData, 2018)*
Professor Luciano Floridi from Oxford University has written extensively about data and information ethics and sits on the EU’s Ethics Advisory Group - he argues that data privacy is critical because it amounts to the protection of personal identity. In others words, he says that we should understand “my data” as we do “my hand” rather than “my car” because it is constitutive of who we are and who we will become. With such high stakes it is likely that breaches, hacks and other cases of misuse will continue to build and build in their significance.

At this time, it is perhaps more obvious to the public that any compromise of healthcare or banking data can be catastrophic - our health and our financial circumstances are typically private matters - but actually companies are now guardians of a whole plethora of different types of information that, in the wrong hands or with the wrong objectives, can be used to construct an uncomfortably intrusive understanding of who we are and how we live. Moreover, when data is collected and analyzed, firms need to be certain that any information, assignations or takeaways from that data is not harmful to data subjects in anyway, either in its sentiment or in its use, particularly when it comes to personal security. (Fiona McEvoy, YouTheData, 2018)

Consumers are becoming more hesitant around the applications of AI and machine learning over fear of how their data is used. A recent study by Genpact revealed that 71% of consumers in a poll didn’t want companies using AI that ‘threatens to infringe on their privacy’. People forget, however, that long before ML was common practice, their data was stored and used by corporations. The root of the concern lies not in the fact that the data is used, but the majority of people are still unclear what it’s actually used for. Individuals are hesitant to hand over their information freely online, but when companies are in charge of our financial information, health records, purchase history, education and such like, the responsibility for safeguarding information falls on them.
WHAT ARE THE RAMIFICATIONS OF BREACHES?

Criminal Actors:
There are always criminal actors looking to weaponize data for their own personal, financial, or political gain. Whether it is personal identity theft or doxing, using data to extort money or maliciously attack businesses, or leveraging data for purposes of political espionage, there are plenty of ways in which breaches can wreak devastation on the public. In the very worst case scenarios, breaches could lead malicious actors to compromise vital infrastructure and/or cause serious harm via remotely controlled technologies.  
(Fiona McEvoy, YouTheData, 2018)

Loss of Data
Over 70% of business who face security breaches fail to reopen within a year of the attack, meaning their ability to collect data is stunted. Companies data can be stolen or lost as a result of a breach, causing systems to fail and consumers to face their personal information falling into undesirable hands.  
(Steve Morgan, Cybersecurity Ventures, 2016).

Loss of Finances
The majority of cyber attacks focus on financial gain. It is forecast that by 2021 cybercrime damages will cost the world $6 trillion. As over 50% of cyberattacks are carried out on small businesses, this no only mean that money will be lost, but entire companies will fold.  
(Steve Morgan, Cybersecurity Ventures, 2016).

How can we avoid breaches in security?
Companies need to take responsibility when it comes to the data they share, and who they share it with. Ultimately, this information is often highly networked with a number of parties involved, and where this results in harm (e.g. a privacy breach) we need to think about how we hold all parties responsible - even where the harm may not be directly attributable to their action (cf. strict liability) - in a way that encourages more careful data handling all along the supply chain. Apple Face ID reportedly stores all of its biometric data it uses on the device itself. Potentially more IoT devices could commit to this to reassure users about privacy, but usually the commercial value is such that they would prefer not to.  
(Fiona McEvoy, YouTheData, 2018)

In enterprise and cloud, security is a top worry. Generally speaking, AI is a number one opportunity for everyone. What is happening here is that both security specialists and maldoers are leveraging the advantages of AI. It is extremely good at well-defined, narrow tasks that need many repeated actions. Like any algorithm, it does not get tired. The advantage is that it does not require writing down all the components of said algorithm manually. This creates an opportunity for even faster growth in attacks.  
(Anton Prokopyev, The World Bank, 2018)
With companies and governments now amassing an unprecedented amount of highly vulnerable data, they clearly have to be more vigilant when considering how they use it and also how they protect it. Businesses and researchers alike are increasingly aware that the security and privacy of their data and users must be kept in check, and that they are accountable for the repercussions of any breaches. Open AI, for example, state that their mission is to build safe AI, and ensure AI’s benefits are as widely and evenly distributed as possible; advancing digital intelligence in the way that is most likely to benefit humanity as a whole. (OpenAI, 2018)

Additionally, Google have stated that they will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment. (Google, 2018)

Whilst these internal policies are integral to the development of safer AI, should we be thinking about universal policies to ensure companies of all sizes abide by the same protocol?

Privacy policies are the primary channel through which companies inform users about their data collection and sharing practices. (Hamza Harkous, EPFL, 2018)

Tractica have forecast that by 2026 the amount spent on the prevention against cybersecurity threats will have increased from $57.9 million in 2017 to $5,385.1 million. (Clint Wheelock, Tractica, 2018)

Whilst it’s integral for companies to be considering their internal processes, to ensure a agreeable standard is met across the board, it may be necessary to enforce blanket policies. Without such regulatory frameworks, data will not only be vulnerable to breaches, but there will be an over saturation of biased systems with sub-par privacy and security standards.

Many companies developing AI systems do so in a black box - a non-transparent method which offers no assurance of the systems security. This can then lead to a lack of accountability, for example ‘profiling people who take part in protests will become increasingly easy, even if they cover their faces. It is not always clear when machine learning algorithms are being used, so harms arising out of the use of AI are hard to detect.’ (Privacy International, 2018)

For these reasons amongst others, governments and boards must ensure that privacy and security laws are applied to AI, and must create new frameworks and regulations where necessary.
WHAT’S CURRENTLY IN PLACE?

There are some regulations already in place, as well as others that are being worked on, such as a bill to create an “emerging tech policy lab” within the General Services Administration (GSA) to advise on government use of technology like AI. “Fundamentally Understanding the Usability and Realistic Evolution of Artificial Intelligence Act of 2017” or the “FUTURE of Artificial Intelligence Act of 2017”. Mariana Acuña of Opaque Studios explained that in their current work, they follow IEEE’s legal recommendations for Policies surrounding Cybersecurity and AI Usage.

AI is present and will continue to be the future, we need to be prepared, we need to start establishing processes, protocols and interactions at the workplace, in our case always putting humans first.
(Maria Acuña, Opaque Studios, 2018)

A growing number of legislators have begun focusing on AI over the last year. As AI based products become increasingly ubiquitous, congressional attention will only increase. We need regulations that are open enough for researchers to continue doing their work, as some of them have already been affected by the “glass box” policy, when most of deep learning models are a complete “black box”. We need to have policies in place that take into account ethics, fairness and security, but this could also hinder AI development. We do need to reimagine norms and institutions, there has to be a balance between research openness, versus personal protection and cybersecurity. The government needs to play a big role here, on the one end it needs to be an ally and not use newer technologies to manipulate legislation, policy implementation, election results, etc. Always putting the rights, safety, and well being of individuals first. We also need to answer the questions who will be liable?

GDPR
The GDPR is one way in which consumers data subjects are being handed back control when it comes to their personal information. Individuals now have the right to be forgotten and the right to access their information.

- Informational imbalance (e.g. you know more about me/what I’ve done/where I’ve been than I can reasonably remember or know myself)
- Anonymity being compromised by the comparative analysis of publicly available datasets. Just a small number of seemingly innocuous data points can reveal highly personal information amount specific individuals.
(Fiona McEvoy, YouTheData, 2018)

The Network and Information Security Directive:
The NIS Directive aims to achieve a high, common level of network and information systems security across the EU. This Directive was transposed into UK law as The Network and Information Systems Regulations 2018 (NIS Regulations) on 10 May 2018. It set up a large web of breach notifications, increasing the amount of information sharing for security in the EU; and the upcoming cybersecurity act, which should lead to a certification scheme for cyber-resilience in Europe.

The e-Privacy directive:
An accompaniment to the GDPR, the e-privacy directive looks specifically at online communications and sate. It is the successor to a current similar directive, commonly called the Cookie Law and works in enhancing GDPR it in light of technological developments (specifically the “Internet of Things”). This updated set of regulations covers cookies, electronic communications (think WhatsApp, Skype, Facebook Messenger), soft opt in and B2B consent.
(Andrea Renda, Centre for European Policy Studies, 2018)
THE JOURNEY TO IMPROVED POLICIES

 ✓ The UK has appointed Roger Taylor, co-founder of healthcare data provider Dr Foster, as the first chair of its new Centre for Data Ethics and Innovation, and started a consultation on the centre’s remit. (Consultation on the Centre for Data Ethics and Innovation, UK Gov, 2018)

 ✓ The Machine Intelligence Research Institute, a research nonprofit studying the mathematical underpinnings of intelligent behavior, are striving to develop formal tools for the clean design and analysis of general-purpose AI systems, with the intent of making such systems safer and more reliable when they are developed.

 ✓ Google have built their AI principles around being built and tested for safety. “We Incorporate privacy design principles. We will incorporate our privacy principles in the development and use of our AI technologies. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.” (AI at Google, Our Principles, 2018)

 ✓ The Future of Humanity Institute FHI houses the Strategic AI Research Centre, a joint Oxford-Cambridge initiative developing strategies and tools to ensure artificial intelligence (AI) remains safe and beneficial. (Future of Humanity Institute, Oxford University, 2018)

 ✓ The Cambridge Center for the Study of Existential Risk: goals are to significantly advance the state of research on AI safety protocol and risk, and to inform industry leaders and policy makers on appropriate strategies and regulations to allow the benefits of AI advances to be safely realised.

 ✓ Partnership on AI have published their guidelines for creating safer AI, stating that they will “seek to ensure that AI technologies benefit and empower as many people as possible”, “educate and listen to the public”, “are committed to open research and dialogue on the ethical, social, economic, and legal implications of AI”, among others. (Tenets, 2018)

 DEEP DIVE: What is Privacy Engineering and Why Does it Matter?

 A field of engineering that is in its infancy, Privacy Engineering will set methodologies and processes for mitigating risks, and address privacy and security concerns. It consists of privacy by design, privacy engineerings manifesto, and privacy engineering. As we have seen many times now it is crucial to have privacy enhancing technology, anonymity and de-identification of data. It is a way to make sure that the use of surveillance, big data collection, and artificial intelligence is ethical. At the end an AI system is only as intelligent, ethical or fair as the individuals programming it. (Mariana Acuña, Opaque Studios, 2018)

 Globally, companies need to be evaluating whether their use of customer or public data is reasonable, and within the parameters of what they might expect given the service provided. Rather than waiting for laws, regulation, or other kinds of public response, ethical firms need to put processes in place. Many already are already working out what these new checks need to be. Ultimately, many involve trying to anticipate the consequences of data use, and how it may affect privacy and security along the line. However, even where the consequences appear minimal, it’s important to consider the issues like respect and human dignity. Are there things that companies shouldn’t do with data simply because they are wrong, regardless of the consequences? Companies shouldn’t wait for an incident or a wave of public disapproval before they make some attempt to figure this out. (Fiona McEvoy, YouTheData, 2018)
4. AI ENHANCING PRIVACY

As the landscape of artificial intelligence grows, as do the concerns attached to its misuse.

How is your data being stored, shared and modified? AI is able to source and analyse huge amounts of data on any individual that may have been shared both knowingly and unknowingly, but what is frequently lost in this conversation is the potential for enhancing both the security, and privacy for individuals.

AI is capable of reducing the number of humans who come into contact with their personal information - for processes such as organising information, storing processing receipts, storing photos and such like, there is no longer any need for a human to review the data (Daniel Castro, 2018). In areas with large amounts of data such as finance or healthcare, there tends to be a large scope for human error because of people having to repeat monotonous tasks - here AI has huge potential to mitigate error. (Abhishek Gupta, Microsoft, 2018)

RECAP

Security: Protecting individuals or organisations from using your personal information against you, e.g. health records, personal finances, social security details.

Privacy: Individuals secluding themselves from the public eye.

Today sensitive data is typically protected using rules and heuristics, controlling access, use, dissemination, and so on. Managing and improving these rules is a complex task and often ineffective. Advanced deep learning models are now able to take into account a larger number of parameters for deciding whether to allow an operation or not, and this has the potential of improving privacy, beyond the current manual rules and hard-coded decisions. (Eli David, DeepInstinct, 2018)

IN PRACTICE

People are often more comfortable having their sensitive information ‘seen’ by machines rather than individuals.

Take Woebot, the personal mental health assistant, for example: Alison Darcy, CEO and Founder explained that patients often don’t want to burden family or friends with their issues, or want to keep them private. This therefore allows individuals to seek help in what they consider to be a safe environment. The patients information is not linked in any identifiable way, ‘it is de-identified and anonymous’ and only sees your activity within the conversation between the user and the bot.
Anticipating risk: DARPA launched the Cyber Grand Challenge: a competition to create automatic defensive systems capable of reasoning about flaws, formulating patches and deploying them on a network in real time. The challenge aimed to use AI as defence mechanisms to find vulnerabilities within systems to serve in anticipating risk, and also to detect risk in context where humans can’t recognise or realise without a huge amount of effort.

Hacker prevention: Official government bodies and companies are often one step behind hackers in adopting new technologies. AI will allow organisations to safeguard themselves by building mature predictive and preventative models. Once we look at AI as a method of augmenting our capabilities, it will improve internal operations. CISOC reported that AI can learn to identify unusual patterns in large volumes of encrypted web traffic and automatically alert security teams to the need for further investigation. “Currently, the lack of trained personnel in this area is an obstacle to enhancing security in many areas”, so AI will eliminate this issue.

Where is AI enhancing privacy and security?

Cyber Defence: Barclays Africa are using AI and machine learning to detect and respond to cyber security threats. AI is able to detect “known-unknown” threats (previously unseen variations of known threats, malware subfamilies, or related new threats) and “unknown-unknown” (net-new malware) threats. (Forbes, 2018)

Hacker prevention: Official government bodies and companies are often one step behind hackers in adopting new technologies. AI will allow organisations to safeguard themselves by building mature predictive and preventative models. Once we look at AI as a method of augmenting our capabilities, it will improve internal operations. CISOC reported that AI can learn to identify unusual patterns in large volumes of encrypted web traffic and automatically alert security teams to the need for further investigation. “Currently, the lack of trained personnel in this area is an obstacle to enhancing security in many areas”, so AI will eliminate this issue.

Anticipating risk: DARPA launched the Cyber Grand Challenge: a competition to create automatic defensive systems capable of reasoning about flaws, formulating patches and deploying them on a network in real time. The challenge aimed to use AI as defence mechanisms to find vulnerabilities within systems to serve in anticipating risk, and also to detect risk in context where humans can’t recognise or realise without a huge amount of effort.

I think we’ll start to see machine learning algorithms that are very difficult for attackers to intentionally fool, but I don’t think we’ll see any security guarantees in the form of mathematical proofs of strong protection claims. We’ll start to see a good set of best practices recommendations for how to make machine learning algorithms fair, when they’re used to make decisions that strongly affect people’s lives (like parole decisions, mortgage applications, etc.). We’ll start to see much stronger privacy guarantees, from techniques like differential privacy, federated learning, and maybe even homomorphic encryption.

(Ian Goodfellow, Google Brain, 2017)

Technology is an “enabler”, and can possibly lead to both positive and negative results. However, there is immense potential for it to lead to better privacy protection. In modern day, AI is being used to keep data safe, for example in finance, your transactions via credit card are constantly being monitored by the credit card company, if the algorithms start seeing abnormal transactions based on your history record of purchase, they’ll preemptively cancel or disable the card until they can contact the card holder to verify if it’s legitimate.

Here AI works as a gatekeeper of sorts. It can also be used to differentiate between authorised users and unauthorised users based on certain inputs like biomedic information (like your thumbprint or irises); all of that recognition is a form of AI. This technology is getting better and better, and is not only less prone to fooling, but can also to recognise when people are trying to fool it. (Jeremy Marvel, NIST, 2018)
WHAT IS REQUIRED TO CREATE HIGH QUALITY VIDEO CONTENT?

Large amounts of training data: This needs to be of high quality, which is why in past examples, they have been trained on US presidential speeches. As the systems improve, however, it’s not hard to imagine being able to use surveillance or other self-uploaded footage on social media as fodder to train these systems.

In another line of research and work that was adopted for malicious use, Deepfakes was able to supplant the face of a targeted individual and replace that in defamatory videos - threatening again the integrity of the identity of an individual.

5. CASE STUDY: IDENTITY FRAUD

WHAT MIGHT THE FUTURE OF IDENTITY FRAUD AND THEFT LOOK LIKE?

Imagine a scenario where there exists a perfect replica of you, encapsulating your physical, biological and online characteristics such that it is indistinguishable from you. What if that entity was able to act on your behalf and execute transactions, authenticate and utilize services that you’re authorized to use, interact with your loved ones and more?

While this is something that seems straight out of science fiction, we’re gradually heading towards a direction, powered by continual advances in deep learning techniques, where we are able to emulate the digital characteristics of individuals with an increasing degree of accuracy. While a whole lot of these issues fall squarely within the intersection of deep learning and cybersecurity, some research advances start off as innocuous ways to advance different paradigms within the field of deep learning but can be easily adapted for malicious uses.

As an example, let’s consider the case of voice and video synthesis: Lyrebird is able to take a voice sample of less than a minute long, admittedly of very high quality, to capture the essence of your voice and then generate audio output saying anything that is provided as an input. There is a lot of potential here to impersonate someone on VoIP and phone calls, perhaps even triggering panic responses impersonating a family member.

Research papers have successfully been able to generate convincing video samples of people speaking text inputted by the researchers.

WHAT IS REQUIRED TO CREATE HIGH QUALITY VIDEO CONTENT?

Large amounts of training data: This needs to be of high quality, which is why in past examples, they have been trained on US presidential speeches. As the systems improve, however, it’s not hard to imagine being able to use surveillance or other self-uploaded footage on social media as fodder to train these systems. In another line of research and work that was adopted for malicious use, Deepfakes was able to supplant the face of a targeted individual and replace that in defamatory videos - threatening again the integrity of the identity of an individual.

ABHISHEK GUPTA: BIOGRAPHY

Abhishek is a Prestige Scholar from McGill University with a degree in Computer Science and Economics. He's a software engineer at Microsoft and an AI Ethics Researcher at McGill University, as well as the lead organiser for the Montreal AI Ethics meetup, a bi-weekly event where members from the community do a deep dive into technical and non-technical aspects of the ethical development of AI. His current interests include reproducibility and accountability as key pillars in building ethical AI. Most recently, he has written a paper on how large-scale data breaches combined with advances in deep learning techniques will lead to the next generation of identity theft and fraud.

As an example, let’s consider the case of voice and video synthesis: Lyrebird is able to take a voice sample of less than a minute long, admittedly of very high quality, to capture the essence of your voice and then generate audio output saying anything that is provided as an input. There is a lot of potential here to impersonate someone on VoIP and phone calls, perhaps even triggering panic responses impersonating a family member.

Research papers have successfully been able to generate convincing video samples of people speaking text inputted by the researchers.

WHAT IS REQUIRED TO CREATE HIGH QUALITY VIDEO CONTENT?

Large amounts of training data: This needs to be of high quality, which is why in past examples, they have been trained on US presidential speeches. As the systems improve, however, it’s not hard to imagine being able to use surveillance or other self-uploaded footage on social media as fodder to train these systems. In another line of research and work that was adopted for malicious use, Deepfakes was able to supplant the face of a targeted individual and replace that in defamatory videos - threatening again the integrity of the identity of an individual.
Identity fraud is now entering a territory where there are many more repositories of unalterable personal data. This is a category of personally identifiable information (PII) that can’t be changed in the event of a security breach or identity fraud instance.

What does this include?

- DNA
- RETINA SCANS
- FINGERPRINTS
- ALL BIOMETRICS

While things like phone numbers, addresses, and even national identity numbers can be changed in the event of an identity theft, with quite a bit of effort, biometrics are immutable and hence their loss triggers irreparable damage. Not only that, but as seen in this recent case, DNA information can also affect people closely related to you which makes this information even more valuable.

Data brokers, entities that operate in the background of the internet and physical marketing world, aggregate data from many sources to create rich profiles of individuals that are then used in enhanced marketing. While there hasn’t been a ton of scrutiny on their ways of operation, what is clear is this:

*With the advent of deep learning methods, the depth and richness of information that can be gathered from the massive hordes of data with these firms will only increase.*

Additionally, there are more and more avenues to gather these attributes, say from IoT devices in a home that can further capture habits and patterns of behaviours of individuals and then use them in perpetrating ever-more convincing identity fraud. **Recent work in Generative Adversarial Networks (GANs) has also enabled the generation of synthetic data** that can be used to further augment the availability of data to train these systems.

All of this not only amplifies the challenges that we face from a cybersecurity perspective but also brings into question how we handle release of data in the public sphere. Take for example, deanonymization and subsequent disclosure of sexual preferences of an individual done by matching up IMDB ratings of rare movies with an open Netflix dataset. Another case in point was the publicly accessible, yet anonymized, dataset released by AOL on search queries that identified a woman living in the US - search queries are often a direct insight into the behaviour, needs and state of mind of an individual and making that data public can have severe consequences. The recent dataset made public by Strava disclosed locations of hidden military bases because of regimented patterns of exercise.

All of the above enhance the amount of data points that one has access to when trying to model an identity, right down to behavioural data, needs and wants. As we inch into the world of the quantified self, we must ensure that ethical and transparent practices are followed, especially when operating on PII and unalterable personal data. There are quite a few resources on good data hygiene, robust machine learning practices that will help to mitigate potential harmful scenarios.
6. CASE STUDY: GOVERNMENT

JEREMY MARVEL: BIOGRAPHY

Jeremy Marvel is the project leader for the Performance of Collaborative Robot Systems, which is developing test methods, metrics, and protocols for the evaluation of human-robot and robot-robot collaboration. Jeremy’s research expertise include intelligent and adaptive solutions for robot applications, with particular attention paid to human-robot and robot-robot collaborations, multirobot coordination, industrial robot safety, machine learning, perception, and automated parameter optimization.

The U.S. National Institute of Standards and Technology (NIST) is a research laboratory for the U.S. Department of Commerce:

Our mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technologies in ways that enhance economic security and improve quality of life.

This project is focused on assembly-based manufacturing, and how robots can contribute to a collaborative operation aimed at completing manufacturing processes.

A large component of what this project works toward is focused on the development of test methods and metrics that can be used to measure the impacts of the use of artificial intelligence (AI) in improving the “collaborative” functions of robot systems. Such functions include collaborative robot safety, the ability to coordinate temporally and spatially with team members, the capacity to communicate high-level task-relevant information, and the understanding of the roles and responsibilities of the respective members of the collaborative team.

This push toward the integration of AI into collaborative technologies has led to some interesting opportunities for which the impact needs to be measured, and the challenges that impede the measurement of effective human-robot interactions that need to be addressed.

Specifically, AI and ML, being non-deterministic (i.e., the results, behaviors, or performance are not necessarily guaranteed), is difficult to definitively verify and validate.

Applying a statistical expectation of performance given sufficient data points can be pursued, but there is never a 100% guarantee that AI will do what you expect or want it to do unless you effectively test into infinity.

Moreover, the quality of any AI and ML approach is wholly dependent on the quality of the data sets used to train the system. Guidance for data set quality is lacking, so it is currently not possible to determine whether a given set of training data will result in good (let alone optimum) performance.
AI-driven applications in manufacturing applications may lack the robustness and reliability necessary for deployment on the factory floor. Moreover, their performance in safety-related software functions cannot be guaranteed, often leading to concerns as to whether they are appropriate as primary protective measures for people in the manufacturing environment. One particular use case we are investigating is the use of AI to drive sensor fusion for human detection and localization for interaction and safety purposes in flexible factory environments.

Neural networks are being trained on raw sensor inputs along with associated metadata to identify whether objects within a scene are classified as human, and where they are located within a complex scene. Here, the algorithms being developed are as prototypes for analysis.

The advancement of the measurement science to evaluate such algorithms and their impact on situation awareness is the principal focus of this effort.

Because of the safety concerns, these algorithms are being tested under controlled conditions, and are used in tandem with (but separate from) the primary safety functions.

Being based on ML methodologies, such approaches pose certain risks in how human data may be captured, encoded, and stored.

- Depending on the contents of the training sets, the recognition of “human” operators may be skewed toward or away from certain postures, ethnicities, genders, body types, facial features, etc.

- This could lead some systems to identify some individuals as “human,” and others as “not human,” introducing unintentional hazards and bias.

- As such, the quality and diversity of data sets becomes imperative.

- Currently there are no standards or established best practices for the assurance of these qualities in a data set containing human subjects.

As such, to attain this, large quantities of data are required. This introduces the potential for certain concerns to arise.

- Data sets may need to be stored long-term or shared to improve performance or retrain safety systems. Storage and access restrictions may be needed to maintain the security of the data, and assure that it cannot be tampered with or altered.

- Many of these data sets are likely to contain the visual likenesses and other physical characteristics of many different individuals, which constitutes personally identifiable information that is thus considered sensitive.

- Privacy challenges thus arise, potentially warranting the encryption and secure storage of the data, and reiterating the necessity for access control, and the approval of an institutional review board for the generation, storage, access, and destruction of the data sets.

- To address these, institutional reviews are required to assess the risks and assure privacy and security concerns are addressed.
7. PREPARING FOR THE FUTURE

“AI-related sensitive data is not and should not be considered different than all other sensitive data. Just as every corporation puts in place an arsenal of cybersecurity solutions to protect against cyber attacks and data leaks, every corporation should protect and guard its AI-related data.”

(Eli David, Deep Instinct, 2018)

Complexity and lack of context are two of the biggest barriers to ensuring future privacy and security in an increasingly AI managed future. Current solutions lack understanding of what they are doing, a fact that is laid beautifully clear when neural networks can recognise a dog as a giraffe merely because some minor tweaks are made to an image that cannot be detected by the human eye.

We will increasingly need to focus on the creation of digital intelligences that are able to adapt to new situations without prior exposure to them, ones that fundamentally avoid the issues around software and data complexity and ones that understand what they are doing rather than employ vast computing and data resources to simply pretend to understand. In security we often use “complexity” and “unknown unknowns” as a form of gallows humour understatement to frame the position we’re in, and in particular, how software autonomy simply doesn’t scale as it is developed now, but the threat is real. A new generation of software autonomy that can genuinely adapt to the scenarios it finds itself in and is actively rewarded or punished for its successes and failures is a potential route through a maze that gets bigger faster than we can figure out ways through it. (Toby Simpson, Fetch.AI, 2018)

AREAS WITH NEED FOR IMPROVEMENT

1) Where there are antiquated systems (like was clearly the case in the Equifax hack) there needs to be an aggressive program of remediation. It isn’t good enough for companies to know that their “security barrier” has major holes and not invest immediately to fix them.

2) Individuals need to play a role in monitoring where they put their data. It is shocking how loose people are with giving out their social security number without really thinking about how it can be used against them. (Catherine Flax, Pefin, 2018)

AVOIDING BREACHES: WHAT’S NEXT?

Right now we witness the beginning of an AI revolution that will completely change how our future is going to look.

We will soon be living in a world where everything is AI-based, from medical diagnosis and treatment, to our autonomous cars, personal assistants, and similar to every advanced technology it will be used for military purposes, surveillance, and also for fraudulent activities as well. In such a world, manipulating AI models could have catastrophic consequences, e.g., erroneous medical diagnosis or treatment, lethal car accidents, etc.

Cyber attacks will not only mean huge financial loss, but in many cases can result in direct life and death consequences. This underscores further the need for better cybersecurity solutions that would protect this imminent AI-based world.

(Eli David, Deep Instinct, 2018)
In the past, anti-malware tools were based on signature detection, the so-called signature-based IDS. Your anti-virus would inform you that it found something if this mechanism was triggered. However, the number of new types of attacks and malware is growing very fast. A lot of the time these are slightly modified versions of existing viruses. Detecting new, unknown signatures the old way becomes inefficient, and AI helps in automating this task.

Right now a typical scenario would be that you set up a new server. Nobody knows its IP address yet. You start getting traffic, operate normally, and collect the data on how it is supposed to operate. Once you have this knowledge base, you could compare new, strange events with your historical evidence. AI is prepared to help you with this, by detecting anomalies in your server's behavior, classifying new events.

Of course, maldoers know about AI-enabled systems as well. And they likely employ it. For instance, using similar technology, they are able to track behavioral trends and patterns of security administrators’ work, how they analyze and monitor their systems. So it becomes an arms race between AI systems. Till this day, human factor remains a key security loophole. Badly designed AI defense? That’s a new vulnerability. The administrator disregarded system warning produced by AI? Here’s another one. In the near future, security systems will adapt to change in the malware landscape on their own. But it is still people who oversee AI, and that’s always been the biggest security concern. (Anton Prokopyev, The World Bank, 2018)

AI is only as smart as the data from which it learns and we are all sources, as well as beneficiaries, of that information. It is imperative that we embrace emerging cryptographic capabilities to train better predictive models without compromising data privacy or integrity. (Jordan Brandt, inpher, 2018)

THE FUTURE OF PRIVACY & SECURITY - WHAT THE EXPERTS SAY:

Obscure AI can be used for blackmail, in gaining access and ‘guessing’ passwords which leads to both personal and security breaches. “Security through obscurity is the worst practice possible, openness and agility is key. We need trusted datasets that are not black boxes, and they need to be adaptable.” With this, we can create a safer future with AI. (Abhishek Gupta, Microsoft, 2018)

Al is only as smart as the data from which it learns and we are all sources, as well as beneficiaries, of that information. It is imperative that we embrace emerging cryptographic capabilities to train better predictive models without compromising data privacy or integrity. (Anton Prokopyev, The World Bank, 2018)

As with advances in any area of technology, new risks and opportunities for exploitation are created. Machine learning adversarial attacks are especially a cause for concern given the potential for widespread use of high-risk classification in fields such as medicine and transportation.

Although these types of attacks have been proven by researchers, they have not yet gained full widespread, mainstream awareness, and the particular systems they target have not been widely implemented. Enterprises should use this time wisely to create risk mitigation strategies before major attacks are perpetrated. (Andrew Clark, Capital One, 2018)
RE•WORK is a female-led events organising company that brings together breakthrough technology, cutting-edge science and entrepreneurship shaping the future of business and society. We create and organise globally renowned summits, workshops and dinners, bringing together the brightest minds in AI from both industry and academia. At each RE•WORK event, we combine the latest technological innovation with real-world applications and practical case studies. Learn from global pioneers and industry experts, and network with CEOs, CTOs, data scientists, engineers and researchers disrupting their industries with AI. We also provide an analysis on current trends and innovations, through podcasts, white papers and video interviews.

Join us to learn more about the most cutting edge AI technologies and how they’re being applied ethically at our upcoming summits. View our calendar of events here.

ADDITIONAL CONTENT FROM RE•WORK

WHITE PAPERS
1. Should You Be Using AI In Your Business? Contributors include OpenAI, DeepMind, Google Research & more
2. The Ethical Implications of AI Contributors include Facebook, Uber, Ethics Research Lab & more

RE•WORK PODCAST: WOMEN IN AI
Women in AI is a biweekly podcast from RE•WORK, meeting with leading female minds in AI, Deep Learning and Machine Learning. We speak with CEOs, CTOs, Data Scientists, Engineers, Researchers and Industry Professionals to learn about their cutting-edge work and advancements, as well as their impact on AI and their place in the industry. Guests include:

- Raia Hadsell, DeepMind
- Negin Nejati, Airbnb
- Laura Douglas, MyLevels
- Georgia Gkioxari, Facebook AI Research
- Jana Kosecka, George Mason University & Google
9. REFERENCES

7. *pribot.org*, Aug 2018
9. *’Data is a Fingerprint’: Why You Aren’t as Anonymous as You Think Online*, Olivia Solon, *The Guardian*, Jul 2018
10. *Tenets*, *Partnership on AI*, Jul 2018
12. *AI offers opportunity to increase privacy for users*, Daniel Castro, Alan McQuinn, *iapp*, Jan 2018
13. *Annual Cybersecurity Report*, *Cisco*, 2018
16. *Cyber Grand Challenge (CGC)*, Mr Dustin Fraze, *DARPA*, 2018
18. *We create the most realistic artificial voices in the world*, Lyerbird, 2018
24. *The Evolution of Fraud: Ethical Implications in the Age of Large-Scale Data Breaches and Widespread Artificial Intelligence Solutions Deployment*, Abhishek Gupta, 2017
27. *Strava’s fitness tracker heat map reveals the location of military bases*, Andrew Liptak, *The Verge*, Jan 2018