



MODERN BRITAIN: GLOBAL LEADER IN ETHICAL AI

YOUNG
FABIANS

YOUNG
FABIANS
TECH NETWORK

Edited by **Marcus Storm**

Forewords by **Darren Jones MP**, Chair of the BEIS Select Committee and the Institute of AI,
Ivana Bartoletti, Chair of the Executive Committee of the Fabian Society

The astounding breadth and detail of AI, coupled with the enthusiasm of the contributors, has turned this into the biggest pamphlet the Fabian Society has ever published. My sincerest thanks to all the fantastic writers who have contributed their talents, whose biographies are published next to their articles.

We are grateful for the donation made by The Coalition for a Digital Economy (Coadec), who have enabled this crucial debate to progress further and wider.

© 2020 Young Fabians

Modern Britain: Global Leader in Ethical AI

Edited by Marcus Storm

www.marcus-storm.com

publications@marcus-storm.com

First published September 2020. Launched at the Labour Party Annual Conference 2020.

ISBN 978-0-7163-2068-5

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher or editor, except in the case of brief quotations embodied in critical reviews and certain other non-commercial uses permitted by copyright law, where the source of information is acknowledged as this publication.

Please send a copy of the document in which this publication is used or quoted to the publisher and editor. For permission requests, write to the publisher or editor, addressed "Attention: Permissions Coordinator".

Like all publications of the Fabian Society, this report represents not the collective views of the Society, nor necessarily the views of the editors nor the writers of the forewords, but only the views of the individual writers. The responsibility of the Society is limited to approving its publications as worthy of consideration within the Labour movement.

Typeset and Cover Design: Robin Wilde: <https://robinwilde.me>. The editor would like to thank Labour Party Graphic Designers for putting him in touch with Robin.

The moral rights of the author have been asserted.

British Library Cataloguing in Publication data. A catalogue record for this book is available from the British Library.

To find out more about the Fabian Society, the Young Fabians, the Fabian Women's Network and our local societies, please visit our website at www.fabians.org.uk

Published by:

Young Fabians

info@youngfabians.org.uk

www.youngfabians.org.uk

Young Fabians

61 Petty France

London, UK, SW1H 9EU

CONTENTS

| | |
|---|----|
| Introduction | 2 |
| Foreword: Darren Jones MP | 3 |
| Foreword: Ivana Bartoletti | 4 |
| Executive Summary | 5 |
| | |
| Section 1: An Explosive Industrial Revolution | |
| Learning from Machines | 7 |
| Money Machines: Artificial Intelligence and Financial Services | 15 |
| The Changing Face of Public Order, and the Digital Self | 19 |
| Tech Frontier: Reprogramming an Algorithmic Patriarchy | 24 |
| Artificial Intelligence in Healthcare: Britain 2020 | 30 |
| Automating Defence: Innovation and the Increasing Role of Artificial Intelligence | 44 |
| AI and Disinformation, an Algorithmic Assault on Democracy | 50 |
| | |
| Section 2: Global Competition | |
| The Artificial Intelligence Landscape | 57 |
| The European Union: Intensifying Competition | 62 |
| America: Birthplace of the Internet | 64 |
| East Asia: Powerful States | 66 |
| AI in Foreign Policy | 70 |
| | |
| Section 3: Conclusions: Policy Proposals | |
| Introduction | 75 |
| Embracing the Potential: Proposals for an AI Regulator | 77 |
| Data Saves Lives - Consider it Vital Infrastructure | 80 |
| Creating a Fair and Competitive Marketplace | 83 |
| Outcomes of a Responsible Regulatory Model | 85 |

MARCUS STORM

INTRODUCTION

Artificial Intelligence is one of the greatest challenges Britain will face this century. It affects everything from education policy, to how scientists research climate change, to our national security.

It is also one of the greatest, most unique opportunities for Britain which may not come again for generations. Levelling up our country and increasing the prosperity and real wages of our citizens relies on bringing our productivity back up to our international peers' levels. Technological adoption has always been the foundation of this process.

The key driver of these new technologies, data, is both infrastructure and a strategic national resource that the Government is not taking seriously enough, both in terms of investment and being forwards-thinking on protecting civil liberties such as privacy.

There is currently a short window for us to plan and adopt a powerful, modernising national industrial strategy which cements us as a long-term global leader in the field.

The lack of clear standards for data and AI, both nationally and globally, hinders adoption and spread in industry and research and without them, the largest markets for AI such as healthcare and financial services are opaque and inefficient. Regulation, if adopted too late and allowed to become too re-

active, can stifle innovation. Proactive, innovative, and engaging regulators, conversely, can be a net asset in helping industries thrive.

A clear legal and ethical framework implemented by a national AI regulator will not only better shape and promote a new, ethical market for the sensitive parts of our economy,



reassure and enable regular companies to adopt productivity-boosting tech, and nudge researchers to identify practical challenges - only a handful of papers last year were published on ways to attack and subvert AI - but will also preserve the explosive and inspiring innovation for other industries.

In Section 3, I introduce the concept of the tripartite social contract, a settlement between humans, the state, and human-like machines. This thinking is critical to determining what kind of future society we want to live in.

For the country, it is a chance to prove our global credentials by be-

ing one of the first countries to take this crucial step. The US and the EU both suffer from fractured data protection rules, and the statist approach of East Asia is not compatible with our liberal democratic values. Genuine global leadership can come from Britain if we foster a reputation for excellence - companies may adapt to our standards first and then export their models to the rest of the world.

Furthermore, economic history, and recent events, remind us that the unidirectional path to more globalisation is not inevitable. A decoupling driven by the seemingly easy removal of several blocks that the global order, and Britain, depends on, leaves many states in a precarious position. We must choose whether to continue to rely on the kindness of strangers or to take a more active role in supporting British economic assets.

Patiently waiting for the private sector to settle on industry standards when private and public sector alike are struggling with the requisite understanding and skills will increase risks and delay adoption and the much needed productivity boost - perhaps long enough for the world to move on, and for the historical chance to shape this vital ecosystem to fall out of Britain's grasp. ■

Marcus Storm
Editor

DARREN JONES MP

FOREWORD

Artificial Intelligence, as a general-purpose technology, will transform the global economy not just in its direct applications, but also in the indirect innovations it will enable. From highly intelligent automated problem solving, speech recognition, speech delivery and computer vision to the impacts AI will have across health, education, agricultural, transport and so much more: it's easy to understand what the global market has been valued to be up to \$3 trillion.

From the changing future of work, to the personalisation and empowerment of public services, the positive potential of AI is immense. However, forward-looking industrial strategy and responsible regulation will be critical to ensure the benefits of AI are felt by all in society, that these technologies do not foster inequalities or discrimi-

nation, and protect human rights. This pamphlet makes a timely contribution to the AI policy debate, adding to the growing conversation on technology policy within the Labour movement.

The UK is not alone in facing the



challenge of ethical AI, but as countries around the world race to prepare for the coming fourth industrial revolution, we have the opportunity to act as a global leader advocating for fairness, accountability, and transparency. We have a

credible base to build on, which I have had the privilege to highlight in my various international roles on AI regulation. As internationalists facing the global advance of automation, we must advocate for ethical policy and regulation in order to bring about justice for all.

While AI is already all around us, the opportunities and challenges will be felt most by coming generations. As such, it is right that their ideas be heard through pamphlets like this, and forums such as the Young Fabians that provide a policy voice for young leaders on these crucial issues.

This pamphlet steps in with bright ideas, put forward by the policy voices of the next generation, illuminating the path forward towards global leadership. 🇬🇧

Darren Jones
Labour MP for Bristol North West

Darren Jones is the MP for Bristol North West and the Chair of the Business, Energy, and Industrial Strategy Select Committee. He is also the Chairman of the Institute of AI (a global network for legislators engaged in AI regulation), a parliamentary lead at the OECD's AI Observatory and the Vice-Chair (International) of the All Party Parliamentary Group on AI.

IVANA BARTOLETTI

FOREWORD

We live in a new technological and data driven era – one which can still amaze us with news of wonderful creations we can get our hands on today or that are just around the corner. It also amazes us for all the wrong reasons, with its power to harm us too. The scandal of Cambridge Analytica and others was a deafening awakening for many, showing how our digital infrastructure has gone so wild, unfettered and unaccountable.

Artificial Intelligence is now progressing at rocket speed thanks to the availability of data and increased computing powers. AI has huge potential, in medicine, cybersecurity and education – and we should leverage them all. But AI carries huge risks too. Automation of inequality, erosion of human

agency and personal autonomy, surveillance and the manipulation of sentiments and ideas – to name a few.



Now it is the time to ensure that technology works for everyone, and that the digital dividends are distributed fairly. This requires ensuring that people, politics and ideas are back in the driving seat.

First, we need to think of our data

as our most valuable collective resource. As this pandemic has shown us, there is nothing more valuable to everyone than one's personal information.

Second, we need a fitness test of current legislation to see whether it is fit for purpose in the AI era. Finally, we need new mechanisms for trusts and transparency, including kitemarks and appropriate redress.

This is not about hindering progress. The opposite - smart regulation - means that we can leverage the full potential of data and tech and do so responsibly as fully-fledged digital citizens. 🇬🇧

Ivana Bartoletti
Chair of the Executive
Committee of the Fabian
Society

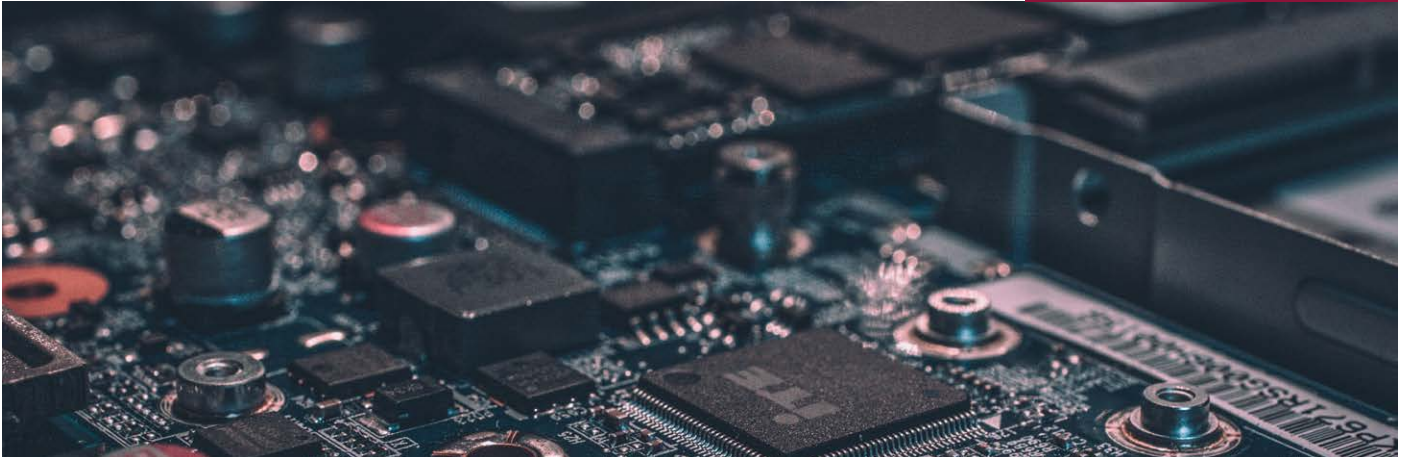
Ivana Bartoletti works, writes, and speaks on privacy and digital ethics, and works with companies in the UK and globally. She is Chair of the Executive Committee of the Fabian Society.

Ivana was awarded 'Woman of the Year' (2019) in the Cyber Security Awards in recognition of her growing reputation as an advocate of equality, privacy and ethics at the heart of tech and AI. She received the Gender Equality Leader Award at CogX 2020.

In May 2018, Ivana launched the Women Leading in AI network, an international lobby group of women advocating for responsible AI.

Ivana is Co-editor of the Fintech Circle's AI Book on how AI is reshaping financial services.

Her first book *An Artificial Revolution: On Power, Politics and AI* is published by The Indigo Press.



EXECUTIVE SUMMARY

Find out about how the incredible speed and impact of AI will affect your sector of interest in **Section 1: The Explosive Industrial Revolution:**

- Learning from Machines explores a future where highly-tailored education is widely available. How will we balance the considerable advantages, particularly for children with special educational needs and disabilities, against privacy concerns?
- Financial Services discusses the ongoing trends in Britain's world-leading financial sector and what we can learn from the early adoption of AI in this field
- Law, Order and Governance explores Home Office issues around face recognition, privacy, and summarises the response to the NHS COVID-19 tracing app
- Equality and Biases discusses the ongoing challenges in ensuring technologies do not entrench existing harms
- Health envisions a future AI-powered healthcare system, with major potential benefits for prevention and care
- Defence and Cyber reveals concerning applications of

AI already used in the old battlefield where drones are replacing soldiers, and in the new battlefield where algorithms battle each other

- Communications discusses the immense impact of fake communications, including the recent usage of GANs to cause major political turmoil in large countries

Research in **Section 2: Global Competition** contains histories and current strategies of three main blocs: the US, the EU, and states in East Asia:

- See State of Play for a global history of AI and a glossary of common industry terms
- America tackles the pre-conception that the US government is more libertarian than ours with regards to state intervention, and what they have done in the past to ensure that their sectors of interest emerge as global leaders
- EU looks at recent successes of the bloc, but also points out limitations and what the UK can do post-Brexit that the EU cannot
- East Asia considers the technologically advanced but more authoritarian states who are using AI in a very dif-

ferent way than we are

- Foreign Policy considers the use of AI as a global political weapon

Section 3: Conclusions: Policy Proposals is the key section for policymakers to consider when drafting or amending bills on data or AI:

- Embracing the Potential: Proposals for an AI Regulator brings together common conclusions from all of the above research to lay out what exactly a UK AI regulator should look like and needs to be to succeed in making the UK a world leader in AI
- Data Saves Lives - Consider it Vital Infrastructure tackles vital considerations about data which are missing from the current Government
- Creating a Fair & Competitive Marketplace summarises existing economics research and commentary and proposes solutions to better the existing environment
- Outcomes of a Responsible Regulatory Model lays out a vision of our future country if all the considerations above are taken into account. 🇬🇧



SECTION 1

AN EXPLOSIVE INDUSTRIAL REVOLUTION

LEARNING FROM MACHINES

By Tom Grand and
Kamal Puwar


Introduction

EdTech companies and researchers are making radical and dramatic changes to learning through the use of AI. We are seeing a wide range of applications and ideas, some of which will have (and are already having) a serious impact on education.

Moving too fast without thought risks causing harm, but moving too slow could lead to missed opportunities and increased inequality. We cannot duck out of this debate - companies and other countries are starting to utilise it heavily, and UK schools are starting to use this technology, without clear overarch-

ing guidance¹. The range of technologies and approaches offers an opportunity to reevaluate and even expand our goals in education, but without clearly defining what we want to achieve we risk letting AI systems and approaches develop that exacerbate the existing problems in our education system.

We approach this by first giving an overview of the current state of play and the directions AI in education is going in - the opportunities and risks, what it looks like for students, and, often less discussed, how AI will affect teaching - particularly from a teacher's perspective. Then

we will discuss three universal issues - data security, surveillance and discrimination/bias. We have primarily focused on schools, but many of these lessons apply to adult education too. We talk a lot about the issues of AI in this paper, because we are cautiously optimistic that there are great opportunities, whilst being realistic that these can only be truly unleashed if we resolve these barriers. We are very much at the start of our national journey of AI in Education, and the UK has an opportunity to define and lead the way in how ethical, effective, and educational AI will look. 

Where we are, and where we could go

More understanding, better learning and happier students?

Improving learning with AI fits into two main categories: understanding learning (e.g. through more effective assessments²) and personalising learning (adaptive instruction, but also broader curriculums and adapting to students' emotional and other needs³). The opportunities they offer could significantly change education for students.

One of the big issues in education is understanding what has been learned. At a government level, we can only glean partial views of people's talents (GCSEs, for example, are a momentary snapshot) - and for schools/teachers there's often a trade-off between spending time measuring learning (to measure how to help), and actually helping improve it⁴. AI offers a way to improve the depth and breadth of understanding of what students actually know. Depth means we can gather more data on an individual student's skills (e.g. using automated systems which both teach and

assess simultaneously in a way a teacher cannot) but also that we can gain more insights by comparing individual students' data to others. This enables us to learn precisely where the student is going wrong and why. Breadth means that we can measure a broader range of skills - including soft skills which have been much harder systematically to assess. UCL Knowledge Lab has developed a system called AIAssess which not only monitors knowledge but metacognitive skills - like how aware they are of their own understanding and how to improve it⁵. Betty Brain, another AI system, trains and tests students on how good at teaching they are⁶.

This offers changes to the classroom, through supporting students with more individualised advice, but also with better measures of soft skills and better ways to see which interventions help the best. This also offers the opportunity to change the GCSE exam - in-

stead of the stressful end of year snapshots we could use longer-term data from the classroom, to achieve a more balanced view of a student. The system would be fairer as it would be harder to game, and would reduce inequalities that come from better exam training in some schools⁷. This opportunity is, in fact, one of the Department for Education's (DFE) 'EdTech Goals'⁸.

Better knowledge of students' current skills enables more personalised learning, the most prominent of which is adaptive instruction. A teacher cannot adapt their rate and style of teaching simultaneously with 32 or more students in a class. However, AI systems like UpLearn in the UK⁹ or SquirrelAI in China¹⁰ are starting to provide personalised adaptive instruction. SquirrelAI runs software where students are taught and given guidance and practice questions, adapting the rate as it tracks student progress - with a teacher monitoring and intervening where the AI system cannot

help any further¹¹. UpLearn¹² has a similar model where students use an AI guided system but get tutor support alongside. Though some remove human elements altogether, like Pearson's Aida - a Calculus app which provides adaptive questions and guidance¹³.

Some studies have suggested that AI-based adaptive learning algorithms - like those above - could improve student learning rate by up to 75% higher test scores a year¹⁴. Although this is still highly debated¹⁵, they are nonetheless promising (and potentially dramatic in its impact over how much students could learn over their whole school career). In the short term, these could be used as interventions by schools with small groups, to support students falling behind, or those with Special Education Needs or Disabilities (SEND), for example. Whilst adult education, these low-cost systems could be highly effective in helping people reskill and change jobs - as platforms like Coursera look to use AI

too (ironically, there could be AI elements, in a course on AI¹⁶)!

There is scope not just to personalise the instruction but also the whole learning experience. A particular benefit of more personalised learning may come to help those with SEND. The pooling of data on how specific students in rare categories learn best could be a huge advantage of AI which is trained on large datasets, likely across the world¹⁷. We are already seeing smaller tweaks like using text to speech to help many students effectively understand exam questions/answers¹⁸. There are wider changes which are becoming available where we can use technology to identify students with SEND better but also find ways to support student well being - as one AI system being used to support mental health being used in UK schools shows (this will be discussed in more in the later section on surveillance)¹⁹.

There are fundamental issues of data security, discrimination, intru-

siveness, and teacher oversight that will be discussed in depth further on, but educationally, there is also the complaint that AI is not the best solution for a lot of the problems it claims to solve. The lack of focus on soft skills in schools is not fixed by adding exams for them, for example. AI could absorb and exacerbate issues in our current system. And all these approaches are not moving at the same pace - standardised tests and more rigid instruction methods have simpler goals, require less data and hence are easier to make AIs for (but may also oversimplify what learning actually means in doing so²⁰). The more idealistic goals of more holistic assessments²¹, genuinely personal learning and freedom of choice require more general AI, more complex data and, likely, experts to work on deep and rich curriculums²². Most likely we will want a mix of approaches, but we won't get that without strong advocates for less rigid approaches and a clear definition of our goals of AI in education²³. 🚩



A tool to remove teachers, or a new tool for teachers?

There are essentially two choices - decreasing the use of teachers (both physically and in decision-making) or giving teachers more time and a new set of tools to use. A lot of the emphasis on AI in education seems to be around the first, especially given the teacher recruitment crisis, "in the 12 months to November 2017... over 50,000 qualified teachers in England left the state sector, equivalent to one in ten teachers leaving the profession"²⁴, but also in order

to cut costs. This would miss out on maximising the potential for AI, not only to improve learning, but to empower teachers more, and even improve retention²⁵.

If it becomes one of the goals of AI in education (or an accidental result) to reduce the need for teachers, it will shape our choices towards approaches that may not benefit either student or teacher. This is not just about removing face time with teachers but reduc-

ing the teacher's role in deciding the curriculum and rate of learning (since AI systems like UpLearn already have these baked in²⁶). Learning analytics, in Higher Education since 2011, has shifted focus on numerical values away from professional judgement - and AI offers even more ways to measure externally, without intervention from teachers, or lecturers²⁷. Even if it is not a goal, teachers could face scrutiny in being asked to explain why they did not follow an

analytic system's standard recommendation or be obstructed from identifying flaws in a system²⁸.

This does not necessarily require a change in the AI systems used - those explicitly aiming to remove teachers are rare - but in the emphasis and who is given control over what. In fact, some of the first beneficiaries of AI could be teachers. Narrow AI systems could be central - as they can be good at replicating many of the routine repetitive tasks teachers do, such as data collection and marking and assessments (Ofqual is starting to look at this for GCSE's, with AI as secondary markers²⁹) decreasing pressure on staff, and in turn, improving teacher retention and freeing up time for human-only tasks.

As shown by the DfE's EdTech strategy (though this rarely mentions AI) their goals are much more about freeing up teacher time - cutting marking time on GCSEs by 20% for example³⁰. Systems like SquirrelAI do not have to be disempowering but, if given to teachers to choose how to use it, it can be used alongside traditional styles.

There are grounds for cautious optimism. AI could improve teacher satisfaction and innovation. We can go even further than providing tools to teachers - and start to ask how we get more teachers and students to lead on the implementation of AI in schools. If we look at the rise of flipped learning or reverse classrooms - for example, watching Youtube video tutorials

at home so students can do more deeper/challenging work in class - empowering teachers is likely to lead to unexpected and valuable examples³¹. Encouraging the sharing of ideas and best practice will be essential. DfE has already introduced an EdTech market place and co-produced with the Chartered College of Teaching a FutureLearn course³² on using technology well in the classroom³³. However, these do not explicitly discuss AI systems, and adoption rates are still relatively low (teachers often end up having to do this sort of CPD in their spare time). Even more exciting is the call for more cross-functional teams which develop AI systems - combining engineers and teachers to design systems together³⁴. 🚩

The future of AI in education?

Data - what is actually happening in classrooms?

Control of data is potentially one of the most urgent and significant areas of AI in education that needs to be addressed.

AI systems often need a lot of data to train on, and even more to continue to monitor their efficacy, including checking for biases. There could be increasing amounts of sensitive personal data being obtained - the combination of emotional wellbeing, learning, what activities students are doing, and many other data points. This would give a lot of power to potential abusers who sought to identify vulnerable students, for example³⁵. This would be a particular risk if data or systems were linked up between public sector services³⁶. Schools and groups like Multi-Academy Trusts and Local Authorities will need clear guidance and robust systems to control who can access what, ensure good use, prevent abuse and secure data from hacks. The guidance is unclear - the DfE's data protection toolkit for schools is still in beta from 2018³⁷. This is despite the DfE pushing for increased cloud usage (and hence the potential for more

centralisation of data) in schools³⁸.

Ensuring standards across institutions is particularly challenging when we have a highly fragmented educational infrastructure and different contexts. Students change between many different institutions within education (primary, secondary, university; state, private) and private companies may be servicing the data too³⁹. A lot of government advice on data usage and security is still difficult to put into practice⁴⁰.

This also suggests a likely expansion of a teacher's safeguarding role and schools' data security roles. Teachers are on the frontline using systems and accessing student data and will be the first port of call to check they are safe, making sure only the correct people can access certain data, and meeting guidelines. Beyond this, we will need them to provide oversight and evaluate their efficacy - lack of understanding will make it hard to spot and challenge systematic errors⁴¹. Their role in communicating information about data gathered and different autonomous systems'

decisions to students, guardians and other stakeholders will be vital in ensuring safe use and the protection of students - many of whom will struggle to understand the seriousness or impact of the information they are sharing⁴².

We will also need to think about the relations between the public and private sector - overly rigid regulation will prevent adoption in schools. But also a lack of clarity could cause lots of schools to not use approaches that could be beneficial. Currently, most AI approaches - like UpLearn who offer "A*-A GUARANTEED or your money back"⁴³ in A-level topics they cover, or Aida a calculus app⁴⁴ - are used mainly outside schools, and require payment to access. These could end up exacerbating educational inequality if schools act too slow, and access to educational AI is essentially defined by who could afford to pay⁴⁵. This is a particular concern given that companies outside of school may not be subject to the same regulatory constraints that school providers have (and if not, we should also be asking - why not?). 🚩

Surveillance Education

A glimpse into a Chinese primary school which is piloting AI in education can give us an insight into one possible future of the classroom.

The Jinhua Xiaoshun school in the Zhejiang province of Eastern China trialled an AI system which required students to wear a metallic headband across their foreheads. These headbands measured the students' concentration by collecting neural data through three electrodes: two behind the ears and one on the forehead. The headbands display an LED light across the top to indicate the 'concentration level' of each student. A red light would signify that a student is deeply focused, whereas a white light would indicate they are 'offline', an oddly dehumanising word used by a student at the school. The neural information is then sent to the teacher to identify who is paying attention, before being forwarded to a group chat for parents. One student claims his parents punished him for 'low attention scores'.

The school has also installed surveillance cameras that monitor how often students yawn or check their phones during class. Their school uniforms contain chips that track their locations. Teachers (and students) in the school speak highly of the improved grades seen as a result of the technology. If such facial recognition and surveillance technology raise the 'attention levels' and grades of students, should we even be concerned with their invasiveness?

French Marxist philosopher Louis Althusser polemically argued that schools transmit and perpetuate capitalist principles (such as competition) and instil subservience to authority⁴⁶. The Jinhua Xiaoshun school epitomises how AI can bring

this capitalistic vision of education to new levels. It is not difficult to see how an AI system which superimposes a coloured rectangle on a student's face, either reading: "ID: 000010, State 1: Focused," or "ID: 000015, State 5: Distracted," can further encourage principles of competition. Encouraging such a competitive atmosphere runs the risk of transposing the educational inequality that we see amongst schools to within classrooms.

Considering Althusser's idea that schools instil subservience to authority, it is not difficult to see how authoritative tendencies in schools can lend to systematic abuse. Hartzog and Selinger write that facial recognition software is intrinsically oppressive and argue that 'the sheer intoxicant of power will tempt overreach, motivate mission creep, and ultimately lead to systematic abuse'⁴⁷.

The image painted of the Jinhua Xiaoshun school may seem both dystopic and distant, but here in the UK, 50,000 students at 150 schools are having their mental health monitored by AI systems to detect self-harm and bullying in 2019⁴⁸. At first glance, this appears as though it is in stark contradistinction to the Jinhua Xiaoshun model, as it emphasises student wellbeing rather than grades. However, the University of Buckingham's report on Ethical AI in Education raises the concern that pastoral AI systems as such could fail to identify urgent safeguarding needs to protect vulnerable students⁴⁹. The report also highlights the possibility that the data gathered by such software (such as an individual's focus level or emotional status) can be used in detrimental ways. They give the example of oppressive states using such data to iden-

tify non-compliant individuals⁵⁰. This poses the question: where is the line between measuring a student's mental health, and correlating it with school performance, changing sets and preemptively finding children with lower mental health as potential non-compliant individuals?

If we were to transpose the 'successful' results from the Chinese model onto education in the UK, there is potential for an increase in grades. However, this might peril other purposes of schools, such as effective pastoral care and relationship-building skills. As Didau and others note, a lot of learning in traditional schools is not explicitly taught, but instead learnt through the modelling of teachers and observation of interactions between the teacher and other students. There are many valuable 'hidden lessons' within schools about conflict resolution and interpersonal relationships which teachers model day in, day out⁵¹.

Enhanced principles of competition and a hyper-focus on grades have both defined and ravaged the modern educational system in Britain. On a macro scale, the perceived importance of being placed on national and regional league tables has encouraged an atmosphere of competition between schools, rather than collaboration. Consequently, this has encouraged middle-class parents to relentlessly find a way to push their children to the best schools, leading to furthered educational inequality by driving down standards at less popular schools. A study by the University of Bristol⁵² also found that league tables punish and reward the wrong schools as they fail to take into account factors such as pupil ethnicity, deprivation and

SEND. When we consider these variables, a fifth of schools see their national league table position change. The resulting competitive atmosphere encourages schools to act as marketing enterprises, placing sole value on the immediate results and sacrificing other vital purposes of schools.

Does AI discriminate?

With a 13.2% attainment gap between the likelihood of white students and BAME students achieving a First Class or Upper Second Class Honours at university⁵⁴, it is crucial to ensure the gap is not further exacerbated at a primary or secondary school level. With this in mind, we must consider whether AI will ameliorate or worsen such inequities.

Facial recognition software works by schematically extracting representations of facial features captured by digital video image, assigning numerical values to these representations, and computationally making comparisons between these values and existing data of previously analysed faces⁵⁵. As a result, such software has the poten-

Conclusion

Introducing surveillance AI, particularly facial recognition technology, into schools has a demonstrable propensity to perpetuate class, gender and racial disparities. It is not unfathomable that technology can ameliorate the discriminatory proclivities of facial recognition software in the future, however it is also clear that we are not close to achieving this. Anrejevic and Selwyn succinctly write, of facial recognition software in schools, 'any 'added value' or gained 'efficiencies' are outweighed by the consequences of automated sorting and classification for students.⁶²

With 4.2 million children living in poverty in 2018 and 2019⁵³ (that is nine in a classroom of thirty), schools must offer more than the minimum GCSEs required for Further Education. Pastoral care and personal edification are essential purposes of schools and can more than often contribute towards educational attainment. There is a dan-

ger to engender racial and gender biases. A US federal study of the most widely used facial recognition algorithms found that the majority appeared to sustain racial bias, as there was a higher rate of misidentification for Asian, Black and Brown faces in comparison to White faces⁵⁶. It is not difficult to imagine the discriminatory consequences algorithmic biases may have for BAME students. Stark warns of the consequences of racial biases in facial recognition software as he writes, 'If human societies were not racist, facial recognition technologies would incline them toward racism; as human societies are often racist, facial recognition exacerbates that animus'⁵⁷.

Andrejevic and Selwyn write that

If AI technology is implemented too quickly without care, we risk systematic errors - like the mis-marking of exams⁶³ or systematic biases, for minority groups, but also for groups where data might be more scarce (e.g. non-neurotypical students may not respond as well to new instructive models which have been built with data primarily focused on neurotypical students.) However, acting too slow could also increase educational inequality as is likely that private schools, and wealthy parents, will obtain access to AI technology to advantage their children, at least academically, which would further widen the attainment gap.


ger that a hyper-focus on grades and ineffectual digitisation of pastoral care may reduce schools to a societal mechanism which perpetuates class disparities and ignores risks to vulnerable students. 🚩

facial recognition technology will perpetuate racialised class hierarchies as such technology foregrounds fixed attributions of students' race, but also gender⁵⁸. From an intersectional lens, dark-skinned women are the most susceptible to discrimination, as Steve Lohr notes, 'the darker the skin, the more errors arise—up to nearly 35 percent for images of darker skinned women'⁵⁹. Buolamwini and Timnit's ground-breaking research found that IBM's facial recognition software had a 0.3% error rate for lighter-skinned males and a 34.7% error rate for dark-skinned females⁶⁰. All other classifiers (Microsoft and Face++) also performed significantly better on lighter-skinned male faces⁶¹. 🚩

A neglected area of research in AI education is its impact on adult education and SEND. We acknowledge that there are advantages of AI education, and there may be even more in the future. However, data and surveillance are significant issues to address as they reflect wider societal issues. It is when we address these barriers that we can begin to embrace AI as a tool to improve education. The implementation of AI education must also occur in a fairer educational system with less disparity between schools. If we are to introduce AI systems, which have many gradients in their efficacy and quality, we must ensure that

there is no better technology for some students and worse for others, as this would further drive inequality in education and society. There must also be clearer guid-

ance, safeguarding and teachers and students must be consulted in the processes of developing and implementing these systems. AI is increasingly becoming an unequiv-

ocal reality, and with the appropriate considerations, there is scope for the UK to lead the way for ethical and effective AI in education. 

Policy recommendations

There are three aspects which need to be balanced when applying AI in an educational context: **safety; equity; and progress, and have made recommendations to help maximise each element.** These are some, but not an exhaustive list of, policy recommendations. We would particularly recommend looking at guidance produced by the Institute for Ethical AI in education and their interim report too.

Safety - making sure data is secure, and students are kept safe from their data being misused.

1. Update the data security toolkit (still in beta from 2018) for schools⁶⁴.
2. Cloud software services: update the list of companies that have filled in a self-certification for security standards (not updated since 2017⁶⁵), look to expand the scope of the questions asked and organisations included. And expand the self-certification checklist to beyond cloud providers to all AI providers for schools.
3. Start talks with teaching bodies, unions, research groups about what safe AI in education looks like, with a goal to form guidance (e.g. a framework) for teachers and schools on how to monitor technology effectively. Furthermore, look into questions about who should own educational data obtained by companies servicing schools and how students and schools are compensated better for the value it brings to companies.
4. Consider measures controlling the use of facial recognition


software in schools until discriminatory consequences are understood better and set out in a fair usage framework.

Equity - making sure all students benefit fairly from AI in education.

1. Encourage more investment in research into using AI to support students with SEND and from known disadvantaged groups.
2. Developing standards for measuring disparity, bias, or discrimination in data-consuming and AI systems, and perhaps a self-certification checklist (this doesn't necessarily need to be compulsory, but could be useful if it was, the main thing is who has and hasn't and what they filled in should be easily accessible - on the EdTech marketplace for example.) This would help schools to make sure AI products are not disadvantaging some pupils more than others.
3. Examining the potential of using AI tutoring systems at schools as a way to provide support to students who are falling behind: for example, interventions for pupil premium kids.
4. Re-evaluating the EdTech marketplace, school procurement processes, and resources for teachers to improve market transparency and deliver the best solutions for students with SEND and from disadvantaged backgrounds.
5. Creating a consistent standard of outcomes across schools, and the ability for schools and researchers to monitor external providers' equality of im-

pact.

Progress - making sure AI in education does help improve learning and student well being.

1. Set up an EdTech hub for England (Wales, Scotland and Northern Ireland all have them) - for greater knowledge sharing, and as part of that, a more definite plan to spread best practice and innovation between schools without increasing workload⁶⁶. Ensure that all four the EdTech hubs have strong links to any national AI regulator or auditor to ensure that there is no overlap of duties and a clean regulatory landscape to foster innovation.
2. To continue work on the 2019 EdTech targets - but also to add more expansive targets on teacher workload (e.g. targets to cut marking in total, not just exams) and equality of outcomes.
3. Investment in research behind using AI in more holistic assessments - i.e. options beyond our current 'end of year' exam heavy model, using data over a longer period of time, or assessments which are more soft skills focused⁶⁷.
4. Increase the focus on 'explanatory learner models' - AI systems that aim to understand how students learn and provide actionable and understandable information for teachers and students, rather than just act as "black box" unexplainable models⁶⁸.
5. Look for ways to encourage companies to involve teachers and students in the design and development of AI systems. 



References

1. The Institute for Ethical AI in Education. 2020. Pp. 22
2. The Institute for Ethical AI in Education. 2020. pp.8 and Lucklin, Rosemary. 2017. pp. 2
3. The Institute for Ethical AI in Education. pp.8 and du Boulay, Benedict. pp. 2903
4. Lucklin, Rosemary. 2017. p.p. 1.
5. Ibid. p.p. 2.
6. du Boulay, Benedict. 2019. p.p. 2903.
7. Lucklin, Rosemary. 2017. p.p. 4.
8. Department for Education. 2019. p.p. 33.
9. Uplearn. n.d.
10. Squirrel AI. n.d.
11. Hao, Karen. 2019.
12. Uplearn. n.d.
13. Pearson. n.d.
14. du Boulay, Benedict. 2019. p.p. 2903.
15. Kitto, Kirsty and Knight, Simon. 2019. p.p. 2857.
16. Rosé, Carolyn P. McLaughlin Elizabeth A. Liu, Ran and Koedinger, Kenneth R. 2019. p.p. 2948.
17. Ibid. pp.
18. Department for Education. 2019. p.p. 10.
19. Rowland, M. 2020.
20. Locklin, Scott. 2020.
21. Lucklin, Rosemary. 2017. p.p. 1-4.
22. The Institute of Ethical AI in Education. 2020. p.p. 6.
23. Ibid. p.p. 22.
24. National Education Union. 2019.
25. The Institute of Ethical AI in Education. 2020. p.p. 6.
26. Ibid. p.p. 6.
27. Williamson, Ben. 2019. p.p. 2800.
28. The Institute of Ethical AI in Education. 2020. p.p. 6.
29. Black, Beth. 2020.
30. Department of Education. 2019. p.p. 33.
31. Education Endowment Foundation. 2017.
32. Chartered College of Teaching. n.d.
33. Department of Education. 2019. p.p. 16.
34. Rosé, Carolyn P. McLaughlin Elizabeth A. Liu, Ran and Koedinger, Kenneth R. 2019. pp. 2955.
35. The Institute of Ethical AI in Education. 2020. p.p. 16.
36. Ibid. p.p. 20.
37. Department for Education. 2018.
38. Department for Education. 2019. p.p. 14.
39. The Institute of Ethical AI in Education. 2020. p.p. 18-20.
40. Kitto, Kirsty and Knight, Simon. 2019. p.p. 2862.
41. The Institute of Ethical AI in Education. 2020. p.p. 18-20
42. Ibid. p.p. 7.
43. Uplearn. n.d.
44. Pearson. n.d.
45. Ibid. p.p. 22.
46. Springer, M. 1995.
47. Hartzog, Woodrow, and Evan Selinger. 2018.
48. Rowland, M. 2020.
49. The Institute for Ethical AI in Education. 2020. p.p.13.
50. The Institute for Ethical AI in Education, 2020. p.p.7.
51. Didau, David. 2019. Pp. 43-46.
52. Leckie, George, and Harvey Goldstein. 2017.
53. Department for Work and Pensions. 2019.
54. Amos, V., and A. Doku. UK, Universities. 2019. Pp. 11.
55. Andrejevic, Mark, and Selwyn, Neil. 2019.
56. Grother, P., Ngan, M., & Hanaoka, K. 2019.
57. Stark, Luke. 2019. p.p.53.
58. Andrejevic, Mark, and Selwyn, Neil. 2019.
59. Lohr, Steve. 2018.
60. Buolamwini, Joy, and Timnit Gebru. 2018. pp.9.
61. Ibid pp.8-9.
62. Andrejevic, Mark, and Selwyn, Neil. 2019.
63. Baynes, Chris. 2018.
64. Department for Education. 2018.
65. Department for Education. 2019.
66. Gibbons, Amy. 2020. n.p
67. Lucklin, Rosemary. 2017. p.p. 1.
68. Rosé, Carolyn P. McLaughlin Elizabeth A. Liu, Ran and Koedinger, Kenneth R. 2019. p.p. 2957.

Bibliography

Amos, V., and A. Doku. UK, Universities. "National Union of Students (NUS)(2019) Black, Asian and minority ethnic student attainment at UK universities.# closing-thegap." (2019).

Andrejevic, Mark, and Selwyn, Neil. "Facial recognition technology in schools: critical questions and concerns." *Learning, Media and Technology* (2019).

Baynes, C., 2018. Government 'deported 7,000 foreign students after falsely accusing them of cheating in English language tests'. [online] Independent. Available at: <<https://www.independent.co.uk/news/uk/politics/home-office-mistakenly-deported-thousands-foreign-students-cheating-language-tests-there-sa-may-a8331906.html>>, accessed 2 May 2020.

Black, Beth. 2020. Exploring the potential use of AI in marking. [online] Ofqual. Available at: <<https://ofqual.blog.gov.uk/2020/01/09/exploring-the-potential-use-of-ai-in-marking/>>, last accessed 31 May 2020.

Buckingham Shum, Simon J. and Lucklin, Rosemary. 2019. "Learning analytics and AI: Politics, pedagogy and practices." *British Journal of Education Technology* 50, No. 6, November: 2785-2793. <https://doi.org/10.1111/bjet.12880>

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, account-*

ability and transparency, 2018.

Chartered College of Teaching. Using Technology in Evidence-Based Teaching and Learning (online course). [online] FutureLearn. Available at: <<https://www.futurelearn.com/courses/technology-teaching-learning>> , last accessed 31 May 2020

Department for Education, 2019. Realising the potential of technology in education: A strategy for education providers and the technology industry. [online] Department for Education. Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/791931/DfE-Education_Technology_Strategy.pdf>, last accessed 24 May 2020.

Department for Education, 2018. Open Beta: Version 1.0. Data protection: a toolkit for schools. [online] Department for Education. Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/747620/Data_Protection_Toolkit_for_Schools_OpenBeta.pdf> accessed 21 May 2020.

Department for Education, 2017. Coud (educational apps) software services and the Data Protection Act. [Online] Department for Education. Available at:

<<https://www.gov.uk/government/publications/cloud-software-services-and-the-data-protection-act>> ac-

cessed 28 June 2020.

Didau, David. Making Kids Cleverer: A manifesto for closing the advantage gap. London:Crown House Publishing Limited, 2016.

du Boulay, Benedict. 2019. "Escape from the Skinner Box: The case for contemporary intelligent learning environments." *British Journal of Education Technology* 50, No. 6. November: 2902–2919. <https://doi.org/10.1111/bjet.12860>

Education Endowment Foundation. 2017. Flipped learning. [online] Education Endowment Foundation. Available at: <<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/flipped-learning/>> last accessed 31 May 2020.

Gibbons, Amy. 2020. Coronavirus: National edtech hub needed to close gap. [online] TES. Available at: <<https://www.tes.com/news/coronavirus-national-edtech-hub-needed-close-gap>> last accessed 28 June 2020.

Grother, P., Ngan, M., & Hanaoka, K. 2019. Face recognition vendor test part 3: <https://doi.org/10.6028/nist.ir.8280>

Hao, Karen. 2019. China has started a grand experiment in AI education. It could reshape how the world learns. [online] MIT Technology Review. Available at: <<https://>



www.technologyreview.com/2019/08/02/131198/china-squirrel-has-started-a-grand-experiment-in-ai-education-it-could-reshape-how-the. last accessed 12 May 2020.

Hartzog, Woodrow, and Evan Selinger. "Facial Recognition Is the Perfect Tool for Oppression." Medium (2018).

Households Below Average Income, Statistics on the number and percentage of people living in low income households for financial years 1994/95 to 2017/18, Tables 4a and 4b. Department for Work and Pensions, 2019.

Kitto, Kirsty and Knight, Simon. 2019. "Practical ethics for building learning analytics." British Journal of Education Technology 50, No. 6. November: 2855–2870. <https://doi:10.1111/bjet.12868>

Leckie, George, and Harvey Goldstein. "The evolution of school league tables in England 1992–2016: Contextual value added, 'expected progress' and 'progress 8'." British Educational Research Journal 43, no. 2 (2017): 193-212.

Locklin, Scott. 2020. "Andreessen-Horowitz craps on 'AI' startups from a great height." [online] scottlocklin. Available at: <https://scottlocklin.wordpress.com/2020/02/21/andreessen-horowitz-craps-on-ai>

startups-from-a-great-height/, accessed 28 April 2020.

Lohr, Steve. "Facial Recognition Is Accurate, If You're a White Guy.", New York Times. Available at: <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>. (2018) last accessed 12 May 2020.

Lucklin, Rosemary. 2017. "Towards artificial intelligence-based assessment systems." Nature Human Behaviour 50, No. 28, February: 1-4.

National Education Union, 2019 Teacher recruitment and retention. [online] National Education Union. Available at: <https://neu.org.uk/policy/teacher-recruitment-and-retention>. last accessed 24 May 2020.

Pearson. Aida. [online] Pearson. Available at: <https://www.pearson.com/us/higher-education/products-services-teaching/learning-engagement-tools/aida.html>. last accessed 31 May 2020.

Rosé, Carolyn P. McLaughlin Elizabeth A. Liu, Ran and Koedinger, Kenneth R. 2019. "Explanatory learner models: Why machine learning (alone) is not the answer" British Journal of Education Technology 50, No. 6. November: 2943–2958. <https://doi:10.1111/bjet.12858>.

Rowland, M., 2020. Artificial Intelligence Being Used

In Schools To Detect Self-Harm And Bullying. [online] Sky News. Available at: <https://news.sky.com/story/artificial-intelligence-being-used-in-schools-to-detect-self-harm-and-bullying-11815865>.

Sprinker, M. (1995). The Legacies of Althusser. Yale French Studies, (88), 201-225. doi:10.2307/2930108.

Stark, Luke. "Facial recognition is the plutonium of AI." XRDS: Crossroads, The ACM Magazine for Students 25, no. 3 (2019).

Squirrel AI. How we're revolutionizing education. [online] Squirrel AI. Available at: <http://squirrelai.com/about>. last accessed 6 June 2020.

The Institute for Ethical AI in Education, 2020. Interim Report. Towards a shared Vision of Ethical AI in Education. Buckingham: The University of Buckingham.

Uplearn. A*-A Guaranteed or your money back [online] Uplearn. Available at: <https://uplearn.co.uk/>. last accessed 21 May 2020.

Williamson, Ben. 2019. "Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education" British Journal of Education Technology 50, No. 6. November: 2794–2809. <https://doi:10.1111/bjet.12849>



Tom is a qualified maths teacher turned software engineer and has worked in a number of EdTech startups. He is interested in all the ways technology can be used to help tackle educational inequality.

Kamal is a Humanities teacher at a secondary school as part of the Teach First Leadership Development Programme. His research interests include educational equality and diversifying the curriculum.

The authors are grateful to Mona Jamil for additional editing assistance for this article.

MONEY MACHINES: ARTIFICIAL INTELLIGENCE AND FINANCIAL SERVICES

By Kyran Schmidt

We are not at Day Zero when it comes to AI in financial services; firms across the financial ecosystem have been experimenting with such technology for years. Adoption is ahead here than other sectors: within financial services, there are strong levels of technological sophistication and a long, if not always storied, history of algorithmic and data-driven thinking. Most importantly, the commercial imperative is real - data is an 'edge' for such firms, so using that data in ever more intelligent ways via AI is a tangible competitive advantage.

Notably, AI is not always just window-dressing within the industry, or confined to a few isolated use cases. There is commonly a degree of exaggeration for marketing purposes, but AI is playing an important role in several areas within financial services. The ability to identify patterns and connections across huge datasets means that AI is proving especially promising in the areas of, for example, fraud detection and risk management. And of course some firms are using the technology for investment management purposes: as early

as 2016, market researcher Preqin estimated that some 1,360 hedge funds (roughly 9% of all funds) made a majority of their financial trades using computer models¹. Naturally, not all these models will be AI-centric, but such techniques will become more prevalent over the coming years with increased commercial and academic attention.

Importantly, AI is not just creating back office efficiencies. AI in fact already touches many UK consumers' lives in two key areas: customer experience and loan decisions. In the UK an significant enabler here has been Open Banking, mandating that banks securely open up their customer data to third-party developers, resulting in richer, more holistic and standardised consumer datasets. Combined with advances in Natural Language Processing (NLP), this has enabled businesses like British startup Cleo which offers automated financial advice via a chatbot interface. In fact, the global growth in 'robo-advisory' apps, offering automated financial advice and investment management, has been marked - some estimates suggest-

ing that assets under (robo)management could hit \$4.1 trillion by 2022². The robots aren't coming; they're already here, and managing people's cash.

Startups like Cleo and other robo-advisor apps are an important part of the picture, but incumbent financial services institutions are also adopting AI at a growing pace. Two thirds of respondents to a study led by the Bank of England claimed to be using Machine Learning in some form, with the most common use cases cited being anti-money laundering (AML) and fraud detection alongside customer engagement. According to this same study, the median insurance firm has 7.5 live machine learning (ML) applications and the median banking firm has 5.5 - while 'the typical firm expects to make, build or deploy close to 20 applications within the next 4 years'³. Financial services firms are both building such applications wholesale themselves and also using third-party software providers to do so, increasing regulatory attention to the ways in which data is stored, protected and shared between different entities. 🇬🇧

The patchwork regulatory landscape

Yet, while AI usage is increasing, the regulatory framework concerning these technologies is still taking shape. In the absence of a standalone, AI-specific regulator, oversight of its use in the sector comes from existing bodies adapting their guidance for this new context. Regulators look to remain principles-based and 'technology agnostic': that is, focused on protecting consumers and preventing abuse, irrespective of the specific technical details.

For the time being then, regulatory direction comes from a patchwork of sources. Sector bodies like the Bank of England and the Financial Conduct Authority naturally play a high-profile role, recently establishing a Public-Private Forum to better understand the current use and impact of AI and assess gaps in current guidance for firms⁴. But for the time being this remains largely consultative, and they have not issued a definitive regulatory framework for AI in financial services; the

EU's General Data Protection Regulation (GDPR), supplemented by the UK Data Protection Act 2018, is in fact the most meaningful piece of legislation when it comes to AI models involving the use of personal data. Alongside this, the UK Information Commissioner's Office - the independent regulatory body charged with enforcing GDPR within the UK - has itself produced a draft framework for firms to audit the compliance of AI solutions with data protection obligations.

Lastly, anti-discrimination law also comes into play, specifically the UK Equality Act 2010 which prohibits

discrimination based on protected characteristics, whether that be via human judgments or algorithmic

decision making. 🚩

A better financial system for consumers - with caveats

The use of AI models in assessing credit risk and making loan decisions brings the importance of anti-discrimination legislation to light, for such techniques risk exacerbating existing unfairnesses.

Entrusting loan decisions to (more) objective, data-driven models has much to commend it, in theory: the role of unconscious, face-to-face human bias in making such decisions is reduced, and firms can process far larger volumes of requests by automating those decisions and incorporating non-traditional data sources into their models. Smaller loans become more economical as a result of reduced human-in-the-loop costs; AI therefore has the ability to boost financial inclusion, compared to traditional methods⁵. It is also worth noting that AI also unlocks policy options - for example, enabling earlier intervention by firms when it comes to detecting vulnerable-looking customers, such as problematic gamblers.

Automation of credit decisions is already happening: in a study led by the World Economic Forum, 38% of respondents in the Deposits and Lending sector reported using AI-enabled credit analytics⁶. One firm at the forefront here is Ant Financial, a spinout of Chinese e-commerce giant Alibaba, which relies heavily on a 'digital core' to 'handle some of the most critical processes and operating decisions' in consumer lending and other financial services products⁷. Ant's 'Zhima Credit', more commonly known as 'Sesame Credit', taps into a user's transactional history in the Alibaba e-commerce ecosystem and combines that with other behavioural indicators to

provide credit scores.

By leveraging non-standard data points in their models, Ant is able to offer loans to segments of the market which are traditionally underserved by other providers. Yet, such offerings can still also unfairly discriminate by favouring those with higher levels of education and social connection in their scoring attributes, to the detriment of financial inclusion⁸. Even with models that do not leverage some of the behavioural indicators which Ant does, there are reasons to worry that increased AI usage may exacerbate bias in consumer lending decisions; for example, a 2019 study by Berkeley academics of discrimination in the American mortgage market found that "at least 6% of Latinx and African-American applications are rejected... [which] would have been accepted had the applicant not been in these minority groups. This amounts to a rejection of 0.74 to 1.3 million creditworthy minority applications."⁹

A big part of the problem here is that the datasets used in such models often crystallise existing biases and patterns of unfairness. For example, if a borrower segment has historically been denied credit for non-commercial reasons (e.g. discrimination on the basis of ethnicity, race, gender or religion), then "the results will be contaminated by the effect of traditionally unfair financial exclusion of those borrowers, and the model will further predict poor credit rating."¹⁰ Apple's credit card sparked controversy in 2019 for seeming to offer smaller credit lines to women than

men, with one person discovering that, despite living and filing joint tax returns with his wife for many years, "Apple's black box algorithm thinks I deserve 20x the credit limit she does."¹¹ There is no intention here to discriminate; it is the indirect result of the data fed into such models. In regulating the use of AI in financial services, much attention must therefore be paid to the types of datasets used and corrective actions taken to ensure non-discrimination.

Fortunately, this seems to be one of the current areas of regulatory attention. The ICO, for example, recommends that organisations "determine and document their approach to bias and discrimination mitigation from the very beginning of any AI application lifecycle, so that the appropriate safeguards and technical measures can be taken into account and put in place during the design and build phase"¹². Statistical methods for measuring levels of algorithmic fairness have also been the focus of research in the past years. We might, for example, insist upon parity in classification levels for different protected attributes – in the credit case, this could mean similar approval or rejection rates for different groups. Yet, consensus on the most appropriate measure of algorithmic fairness has not been achieved; different measures are not always reconcilable. As a Google research piece notes, "mathematics alone is unlikely to lead to the best solutions"¹³; what is most important is choosing some sort of measure and justifying its relevance to the specific case. 🚩

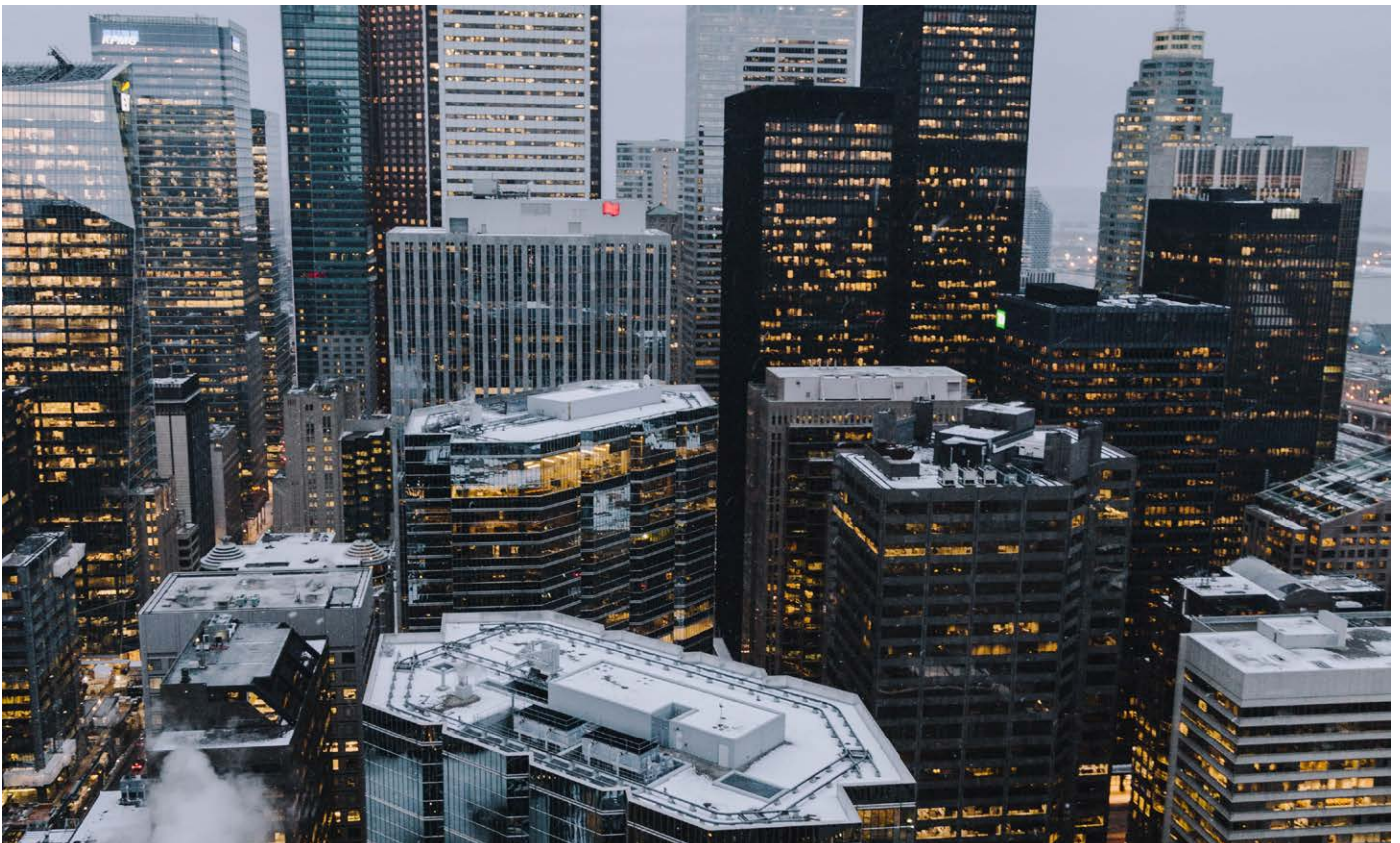
We can safeguard progress

In human-led decision making, or even some forms of algorithmic decision making, we can at least have recourse to explanation. Decisions can be rationalised, rightly or wrongly, by those taking the decision; a borrower was denied a loan because they failed to meet a specific criterion. Some modern AI techniques, such as deep learning, are less amenable to such explanations - decisions are the product of a 'black box', whose mechanisms and logic cannot be easily explained in terms we understand. If we are to place trust in such models then, we must place trust in what goes into them, who produces them and how they are monitored over time; in effect, we require auditability chains for AI use cases.

While we cannot always explain the precise mechanics by which an AI agent arrived at a specific decision, what we can explain is their input (i.e. datasets) and steps taken to ensure such input is fair and appropriate. In a regulatory

context, we should therefore insist on increased transparency in algorithmic decision making – especially so when such services directly touch the consumer. Financial services firms should document their technical architectures and, before deploying models, justify datasets used and steps taken to mitigate against bias. In the same way financial services firms are 'stress tested' - subject to scenario analysis for certain macroeconomic stresses such as drops in GDP or changes in interest rates - they should also be asked to 'equity test' certain decision making models. This could have two parts: first, ensuring that the datasets upon which a model is based are equitable and appropriate; and secondly, simulation testing to confirm that certain attributes do not result in unfair pricing or denial of financial products. To what kinds of decisions these tests should apply, and at what scale, should be a subject of ongoing debate, and regulators should be upskilled to tackle them.

Likewise, we should be paying increased attention to those who produce such models and the governance structures surrounding them; regulatory regimes should further increase in scope to cover the technical professionals responsible for building such models. As the use of algorithmic decision making increases, the boundaries between IT professionals (formerly considered mid-office staff) and managers is becoming fuzzier. The Senior Managers and Certification Regime (SMCR) is a set of regulations introduced in 2016 focused on corporate accountability; it includes provisions for those who perform key roles ('senior management') as well as those who could 'cause significant harm to the firm or its customers', and therefore need to be certified. The technical staff who build and deploy AI models at firms, such as developers and data scientists, should be increasingly subject to the same or a similar certification in light of the increased importance of their work. 🚩





References

1. Cade Metz, "The Rise of the Artificially Intelligent Hedge Fund", WIRED, 2016. Accessed: 18.07.20, <https://www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund>
2. Michael Larner, "Fintech Futures: Leading Innovators, Segment Analysis & Regional Readiness 2019-2024", Juniper Research, 2019.
3. "Machine learning in UK financial services", Bank of England, 2019.
4. "Financial Services AI Public Private Forum", Financial Conduct Authority, 2020. Accessed: 18.07.20, <https://www.fca.org.uk/news/news-stories/financial-services-ai-public-private-forum>
5. Majid Bazarbash, "FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk", International Monetary Fund, 2019.
6. "Transforming Paradigms: A Global AI in Financial Services Survey", World Economic Forum and Cambridge Centre for Alternative Finance (CCAF) at the University of Cambridge Judge Business School, 2020.
7. Marco Iansiti and Karim R. Lakhani, "Competing in the Age of AI", Harvard Business Review 98, no. 1 (January–February 2020): 60–67.
8. Asli Demircug-Kunt, Leora Klapper, and Dorothe Singer, "Financial Inclusion and Inclusive Growth: A Review of Recent Empirical Evidence", World Bank, Policy Research Working Paper; No. 8040, 2017.
9. Robert Bartlett, Adair Morse, Richard Stanton and Nancy Wallace, "Consumer-Lending Discrimination in the FinTech Era", The National Bureau of Economic Research, NBER Working Paper; No. 25943, 2019.
10. Majid Bazarbash, "FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk", International Monetary Fund, 2019.
11. Taylor Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs. The Washington Post, 2019, Accessed: 18.07.20, <https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs>
12. Reuben Binns and Valeria Gallo, "Human bias and discrimination in AI systems", Information Commissioner's Office, 2019, Accessed: 18.07.20, <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/>
13. Martin Wattenberg, Fernanda Viégas, and Moritz Hardt, "Attacking discrimination with smarter machine learning", Google, 2016, Accessed: 18.07.20, <http://research.google.com/bigpicture/attacking-discrimination-in-ml>

Bibliography

"Machine learning in UK financial services", Bank of England, 2019.

Robert Bartlett, Adair Morse, Richard Stanton and Nancy Wallace, "Consumer-Lending Discrimination in the FinTech Era", The National Bureau of Economic Research, NBER Working Paper; No. 25943, 2019.

Majid Bazarbash, "FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk", International Monetary Fund, 2019.

Reuben Binns and Valeria Gallo, "Human bias and discrimination in AI systems", Information Commissioner's Office, 2019, Accessed: 18.07.20, <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/>

Asli Demircug-Kunt, Leora Klapper, and Dorothe Singer, "Financial Inclusion and Inclusive Growth: A Review

of Recent Empirical Evidence", World Bank, Policy Research Working Paper; No. 8040, 2017.

"Financial Services AI Public Private Forum", Financial Conduct Authority, 2020. Accessed: 18.07.20, <https://www.fca.org.uk/news/news-stories/financial-services-ai-public-private-forum>

Marco Iansiti and Karim R. Lakhani, "Competing in the Age of AI", Harvard Business Review 98, no. 1 (January–February 2020): 60–67.

Michael Larner, "Fintech Futures: Leading Innovators, Segment Analysis & Regional Readiness 2019-2024", Juniper Research, 2019.

Cade Metz, "The Rise of the Artificially Intelligent Hedge Fund", WIRED, 2016. Accessed: 18.07.20, <https://www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund>

Taylor Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs. The Washington Post, 2019, Accessed: 18.07.20, <https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs>

Martin Wattenberg, Fernanda Viégas, and Moritz Hardt, "Attacking discrimination with smarter machine learning", Google, 2016, Accessed: 18.07.20, <http://research.google.com/bigpicture/attacking-discrimination-in-ml>

"Transforming Paradigms: A Global AI in Financial Services Survey", World Economic Forum and Cambridge Centre for Alternative Finance (CCAF) at the University of Cambridge Judge Business School, 2020.



Kyran Schmidt is an Associate at Seedcamp, an early stage venture capital firm, where he invests in and supports technology-focused startups across Europe. He holds a first class degree in Philosophy, Politics and Economics from the University of Oxford and is interested in policy issues at the intersection of ethics, business and technology.

THE CHANGING FACE OF PUBLIC ORDER, AND THE DIGITAL SELF

By Anita M. Chandran

“There are no private lives. This is an important aspect of modern life. One of the biggest transformations we have seen in our society is the diminution of the sphere of the private. We must reasonably now all regard the fact that there are no secrets and nothing is private. Everything is public.”

– Philip K. Dick

The increasing presence of AI within our society is already having a significant impact on governance, law, and order. AI-enabled tools for analysing sets of data give the state power to make detailed observations about hitherto unreachable aspects of human life such as migration, the clustering of crime and the spread of infectious disease. However, whenever states are granted new powers, one must ask: how will the state wield this new power? And, perhaps more importantly, how will this affect my ability to be free?

Many are concerned that the state

will have unregulated access to our personal information. Even with the introduction of European GDPR regulations, it is difficult to keep track of our own digital footprints. This leads to a fear of being digitally reconstructed or profiled using our personal data. In modern society, there is not simply a ‘real self’, but also a ‘digital self’: a digital self whom the state and other entities may have complete access to if not controlled and regulated.

The fear of the state monitoring our data leads to a desire to remain ‘off-grid’. People are afraid of sharing their daily internet browsing

history, giving data to private companies such as Apple and Google, and even being recorded by their own mobile phones. This fear can lead to the usage of proxy servers, virtual private networks (VPNs), and encrypted communications. It contributes to a heightened sense of unease. The situation is reminiscent of a quote by American whistleblower, Edward Snowden: “Under observation, we act less free, which means we effectively are less free”. 🚩

Concerns in the age of AI

Considering the nature of data as a resource, the important questions are not about access to our personal data. The state and other actors already have a wealth of personal and anonymised information about the public, from CCTV camera footage, to communications metadata, to patient surveys conducted at GP practices.

The focus is instead on what insights can be drawn from the data. AI algorithms can analyse complex patterns within large datasets and draw conclusions that are not always carefully scrutinised. Furthermore, the very process of drawing these insights raises interesting questions about data consent.

Take the example of a CCTV camera. If we are told that a CCTV camera is installed to monitor a storefront after a crime was committed,

to attempt to catch a perpetrator, to give security to the store owner, and to deter future crime, most people would likely consent. People value their safety and the safety of their society.

If we were instead told that the CCTV camera would assign each person a unique identity which could be detected at the store every time we went, and that this could be used to understand our behaviour as a consumer, it is unclear whether people would continue to consent.

In both cases, the personal data available to the owner of the CCTV camera is the same. So while people are right to be discerning and hard-nosed when consenting to give their personal data away, they must also be wary of the ways in which machine learning and arti-

cial intelligence are used to draw personal insights from that data.

This paper seeks to clarify some of the impacts of government-wielded AI on crime reduction, privacy, and justice. Crime reduction techniques such as predictive policing may have benefits in terms of efficiency and resource allocation but are vulnerable to institutional bias and misuse. Location monitoring applications may aid contact tracing in pandemics such as COVID-19 but may also require citizens to give permissions in an emergency which may be difficult to revoke later.

Lastly, it is vital that a clear, regulatory framework is devised with respect to the usage of personal data and anonymised metadata by the government. 🚩

Predictive policing and crime reduction

There is a certain political appeal to the phrase ‘preventing crime before it is committed’. It is reminiscent of Tony Blair’s 1993 slogan: ‘Tough on crime, tough on the causes of crime’. The prevention of crime before it is committed is the central goal of predictive policing.

Predictive policing uses data analysis to predict and prevent future crime, based on probabilities calculated from past crime data¹. Predictive policing software uses AI to determine individuals at risk of committing crimes, as well as the locations and timings where crime

is likely to occur². This allows law enforcement to efficiently utilise their resources, pre-empting crime and preventing criminal behaviour.

Predictive policing works in two ways. The first is by analysing crime data to determine the hot-spots where crime is likely to occur and at what time. Machine learning software searches for patterns and correlations in large crime datasets. A map is drawn, allowing police to respond early.

The second method in standard predictive policing algorithms is

to identify potential offenders of crime. To do this, the social networks of past offenders are analysed to produce a list of ‘at-risk’ individuals. These individuals can then be monitored, offered home visits from the police, or directed to support services.

At least 14 UK police forces are already attempting to implement predictive policing³. However, implementation before sufficient regulation and consideration can be harmful and divisive. 🚩

Bias and scrutiny in predictive policing

Predictive policing algorithms are susceptible to bias⁴. Their quality depends on the datasets fed into them, and any bias present within them. Take the example of stop and search datasets. Young Black men are disproportionately more likely to be stopped and searched than their white counterparts⁵. This is down to police bias, which is then reflected in datasets. This biased data is then used to train the algorithms which determine the outcome of predictive policing. Proper scrutiny of predictive policing algorithms is vital for reducing bias. If this does not happen, formally-defined mathematical and algorithmic bias turns into racism and discrimination.

However, scrutinising the insights of algorithms may be difficult. Gov-

ernment contracts are frequently outsourced to external contractors, sometimes using proprietary software, which is closed source or privately owned. Therefore, it is almost impossible to scrutinise either by scientific experts or by the public.

As an example, consider the software underpinning Microsoft Windows, or Apple iOS. These pieces of software cannot be copied, modified, or shared. All modifications and updates must be carried out by the proprietor. This makes it difficult to have full transparency on the ways in which these algorithms draw insights and conclusions from large datasets. It is important that software can be peer-reviewed, viewed and discussed by the public or a representative of the public

interest.

Moreover, the general public’s lack of technical literacy in artificial intelligence means that predictive policing algorithms are likely to be seen as ‘black-boxes’. In other words, they are ‘trustworthy, government-certified’ programmes which use ‘computational algorithms’ to deliver accurate results. This lends legitimacy to any conclusions drawn by such algorithms, even though those conclusions may be factually incorrect, biased or improperly scrutinised. Increasing technical literacy of the regulators who represent the public, as well as making sure they have the remit to properly audit and probe these black-boxes, can help promote safe adoption and usage. 🚩

The consequences of predictive policing

Predictive policing can have other negative consequences. Individuals on a list of ‘potential perpetrators’ may be profiled or targeted unnecessarily. Moreover, police are more likely to respond aggressively when expecting violence, which may lead to presumptive police force against innocent


people. When the individuals on perpetrator lists are likely to be from Black and ethnic minority backgrounds due to biased input data, this leads to innocent BAME individuals being more at risk of harm.

Moreover, preventative policing

is a strategy which places police at the heart of crime prevention when it is likely that other methods of community support would be equally or more effective⁶. Alternatives to policing may also improve between disenfranchised communities, communities of colour, and the state.

Fundamentally, predictive policing changes the attitude of policing within the state, from a reactionary

force to a proactive one: a change in 'attitude' which may amount to police departments increasing

their use of cautionary arrests and racial profiling⁷. 

COVID-19 and contact tracing software

Recently, with the spread of a new viral pandemic, a new term has made its way into the sphere of public discourse: 'contact tracing'. Contact tracing refers to the Government's ability to track the movements of infected individuals and reduce the spread of infectious disease.

Contact tracing can be done through contact tracing applications ('apps') where citizens give up data such as location (which may remain anonymised) to the Government. In exchange, they receive information about their level of risk of contracting the disease. This can take the form of increased freedom of movement.

In China, as the COVID-19 lockdown ended, one such contact tracing app was released to some effect. This app works on a 'traffic light' system, analysing location data and coronavirus diagnosis data from residents' smartphones. Each person is then given a colour code which determines their level of contagion risk and dictates their level of freedom of movement. The app is easily installed and used in over 200 cities in China by millions of people⁸.

In cities such as Hangzhou, it is now difficult to travel without showing a colour code. This means there is a level of soft pressure placed on citizens to consent to using the contact tracing app. While this increases user buy-in and potential efficacy of the tracing scheme, it may make some citizens feel uneasy.

As the tracing software is proprietary, the owner company has no obligation to be transparent about how the data is used to arrive at an individual's level of risk. When a reward such as a degree of freedom of movement is assigned to

these classifications, this has the potential to slip into difficult territory: seemingly healthy people being denied freedom of movement for reasons which are totally opaque to them. See the section on China on page _ to find out more about Chinese deployments of AI.

The UK version of a contact tracing app, developed by NHSX, will require high uptake to be effective and for 80% of the contacts of index COVID-19 cases to be contacted⁹. Nominally, all data gathered will be anonymised, and used for health and research purposes. To be effective, individuals will have to opt into record their symptoms. This requires a large degree of trust in the government and their handling of personal data.

Initial versions of the app used a centralised model, in which contact tracing and subsequent analysis is done by a central government server (and not on individual smartphones). This raised concerns about data anonymity, with privacy researchers such as Dr Yves-Alexandre de Montjoye (Imperial College London) outlining possible ways for individuals to be identified from centralised data¹⁰. Even though other approaches are now being investigated, this has impacted public trust.

One of the main motivators for having a centralised contact tracing system is so that the government can better understand the spread of COVID-19. The government can verbally assure the public that all due diligence will be done when it comes to data handling, claim that personal information will be deleted, or that secondary usages of the data will be carried out with stringent regulation. However, without transparency, these reassurances become functionally meaningless¹¹.

These privacy concerns may become a barrier to uptake and usage of contact tracing apps. This has been the case for similar centralised contact tracing apps such as TraceTogether in Singapore (with 25% uptake by population) and smittestopp in Norway (with 20% uptake in the adult population)^{12 13}. If uptake remains low, then contact tracing apps simply will not be effective. To increase efficacy, governments may make the use of contact tracing apps compulsory or make their installation a condition of returning to work.

Much of the UK population might agree with using contact tracing apps as a way of reducing the effects of the COVID-19. However, once again, we see the potential for huge datasets to be given over to the government with limited accountability. It is unclear what later analysis may be carried out on this data, or how it will be used to make inferences about our individual health and wellbeing. Even if the public consents to data being given away in an emergency for our protection, would we still consent if that data were used to form opinions about our personal lives?

A welcome step taken by the UK government is publishing the open source code behind the NHS COVID-19 app. This will enable scientists, software engineers and the public to access and review the app functionality, enabling better scrutiny of the contact tracing software.

It is worth noting that there are other reasons why a lack of scrutiny may occur. Firstly, due to a lack of technical literacy in the press and in the public. This may result in concerns about the app being poorly communicated from scientists and software developers. It

also means that non-scientific concerns about the app may go amiss in the discourse.

The sense of urgency afforded by a massive global health crisis also enables technology to be rushed

The need for a robust, ethical framework

Large datasets taken from the public can be used by governments to inform law, order, and governance. While there are many potential benefits to using machine learning to aid societal governance, it is also vital that the rights of citizens are protected.

In Europe, the necessity of ethical frameworks in artificial intelligence is clear. The European Commission for the Efficiency of Justice (CEPEJ) has already adopted an ethical charter on the use of AI with reference to judicial systems.

Policy recommendations

The advent of AI creates huge opportunities for development in our society, particularly in the fields of law, order, and governance. Though new technology will become available to us, the effect of uptake, the ethics of data storage and the maintenance of public trust are vital.

In the case of policing and police-related algorithms, data can be badly utilised or suffer from inherent bias. This can lead to further bias in outcomes of predictive policing. At the very worst, this leaves vulnera-

ble citizens at risk, and leads to racial profiling. Furthermore, simply trusting the outcomes of predictive policing algorithms can lead to lack of scrutiny and wasted resources. Lastly, a reliance on predictive algorithms to inform policing may put police at the forefront of crime prevention, where other services such as community support may be better placed.

into public usage through necessity, without full consideration of loopholes. People are also likely to sacrifice privacy for 'the greater societal good'. These sacrifices can be buried in the face of more

This framework guides legislators, justice professionals and policy makers when considering the rapid development of AI in the judicial system. The CEPEJ ethical charter is based on several core principles: from respecting fundamental rights; to preventing discrimination; to transparency, impartiality, and fairness¹⁵.

The need for robust ethical frameworks surrounding artificial intelligence also exist in the UK. The public are uneasy about AI and the ways in which their data will

be used to govern them. Even the most well-intentioned Government schemes to use these new technologies can fall foul of exploitation and structural bias.

Where it comes to COVID-19, the Government and NHSX have established an Ethics Advisory Board focusing specifically on COVID-19 app data. More broadly within the UK, the Centre for Data Ethics and Innovation is an independent advisory board aimed at developing governance for AI and data-driven technology.

pressing current affairs, allowing the government to keep expanded powers for much longer than anticipated. 🇬🇧

and lack of user buy-in. This can make contact tracing apps ineffective. When buy-in does occur, it may leave citizens vulnerable to having their data analysed in ways which they do not understand they have consented to.

It is vital that clear and transparent frameworks for protecting individual rights are put into place, and that the government is scrutinised for its use of AI in its governance of the country. 🇬🇧

It is vital that clear and transparent frameworks for protecting individual rights are put into place, and that the government is scrutinised for its use of AI in its governance of the country. 🇬🇧

and lack of user buy-in. This can make contact tracing apps ineffective. When buy-in does occur, it may leave citizens vulnerable to having their data analysed in ways which they do not understand they have consented to.

References/Bibliography

1. Meijer, A. and Wessels, M. "Predictive policing: Review of benefits and drawbacks." *International Journal of Public Administration* 42.12 (2019): 1031-1039.
2. Ratcliffe, J. H. "The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction." *Police practice and research* 5.1 (2004): 5-23.
3. Couchman, H. "Policing by Machine", *Liberty* (2019) <<https://www.libertyhumanrights.org.uk/issue/policing-by-machine/>>
4. Perry, W. L. et al. "Predictive Policing", *The RAND Corporation, Safety and Justice Program* (2013)
5. Babuta, A. and Oswald, M. "Data analytics and algorithmic bias in policing", *Briefing Paper, Royal United Services Institute for Defense and Security Studies* (2019) <https://rusi.org/sites/default/files/20190916_data_analytics_and_algorithmic_bias_in_policing_web.pdf>
6. Kutnowski, M., "The Ethical Dangers and Merits of Predictive Policing", *Social Innovation Narratives, Journal of CSWB, VOLUME 2, NUMBER 1* (2017) <<https://journalcswb.ca/index.php/cswb/article/view/36/75>>
7. "Ethics Advisory Report for West Midlands Police", *The Alan Turing Institute* (2017) <https://www.turing.ac.uk/sites/default/files/2018-11/turing_idepp_ethics_advisory_report_to_wmp.pdf>
8. Mozur, P., Zhong R. and Krolik, A., "In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags", *New York Times* (2020)
9. Meeting notes, Thirty-second SAGE meeting on COVID-19 (2020) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/888807/S0402_Thirty-second_SAGE_meeting_on_Covid-19_.pdf>
10. Radaelli, L. et al., "Quantifying Surveillance in the Networked Age: Node-based Intrusions and Group Privacy", *arXiv preprint* (2018)
11. Stokel-Walker, C., "Coronavirus has ushered in the terrifying era of coughing assaults", 2020 <<https://www.wired.co.uk/article/coronavirus-coughing-spitting-assaults>>
12. Abbas, R. and Michael, K., "The coronavirus contact tracing app won't log your location, but it will reveal who you hang out with", *The Conversation* (2020) <<https://theconversation.com/the-coronavirus-contact-tracing-app-wont-log-your-location-but-it-will-reveal-who-you-hang-out-with-136387>>
13. [13] Kerr, B., "One in five shares data from Smittestopp", *Norway Today* (2020) <<https://norwaytoday.info/news/one-in-five-shares-data-from-smittestopp/>>
14. [14] Findlay, S., Palma, S. and Milne, R., "Coronavirus contact-tracing apps struggle to make an impact", *Financial Times* (2020) <<https://www.ft.com/content/21e438a6-32f2-43b9-b843-61b819a427aa>>
15. [15] CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment (2019)



Anita Chandran is a PhD student in Physics at Imperial College London and a writer whose work focuses on AI, science and historical fiction. She is the fiction editor of Tamarind Literary Magazine.

Twitter: @tintachan

Instagram: @anitamchandran, @tamarindlitmag

FIGHTING ENDEMIC BIASES ON THE TECH FRONTIER: REPROGRAMMING AN ALGORITHMIC PATRIARCHY

By Cecilia Eve

THE PRESENCE OF AI IN YOUR LIFE: CLOSER THAN YOU REALISE

Artificial Intelligence, to the average person, invokes images of a sentient computer overlord from science fiction, or David Hanson's social humanoid robot Sophia whom Piers Morgan asked on a date (and was satisfyingly rejected by).

But it couldn't be further from its remote, mysterious and largely irrelevant reputation. In the modern world, AI is an integral part of your day. Ordered an Uber, or UberEats? It uses AI to predict your ETA. Stuck in traffic? Your maps app or live SatNav will be using AI to produce more efficient routes for you. Googled something? Google's predictive analytics will be working to generate personally tailored re-

sults. Asked Siri or Alexa anything? They both use cloud-based AI technology. The UK is one of the Digital 9; a global network of countries which use predictive algorithms in decision-making, at governmental level and many others. The reality is, AI is already an integral part of our everyday lives.

AI, and generally, automation, has long been considered by companies as a mechanism to improve processes from hiring, to supply chains, to using Big Data to produce data driven conclusions that regular management tools struggle to even capture. If you apply for a job, your candidacy may well be assessed by a neural network (systems which loosely model the

human brain) rather than a human, and you're very unlikely ever to know. Companies also see AI as the solution to longstanding systemic biases; a machine learning algorithm surely cannot have an agenda, right? The inductive logic here seems reasonable. And yet more and more within the industry have expressed growing concern with its implementation. Instead of eradicating biases, research is showing that in some applications of AI, it is instead perpetuating them. Here we look at the real world applicability of AI, how it may be further entrenching existent social inequalities, and how we can address these issues. 🇬🇧

AI hiring automation: downgrading women's CVs

An increasing list of very familiar companies are automating their hiring processes with AI programs, including Goldman Sachs Group, Hilton Worldwide Holdings, Netflix, United, Cisco, and Cognizant.

Their rationale is that by automating this process, they will reduce inefficiency, increase productivity, and save money. Not only that, but they can save huge amounts of time and turn gathered data into actionable insights at a level which has yet to be reached by human workers. LinkedIn, the world's largest professional network, has gone further. It algorithmically ranks candidates, assessing their fit for available vacancies on the website. Entelo is a world leading AI-driven recruiting automation platform, and according to their report¹ surveying hundreds of leaders in talent acquisition, 61% of

recruiters believe that automation will also help to eliminate bias and nepotism, whilst instead promoting a more meritocratic format. But the data on the existence of bias does not concur.

Recruitment is one of the most significant arbiters of equal gender representation in society. It informs the closing or expansion of the gender pay gap, making intrinsic bias an inexpressibly serious problem. In 2014, Amazon started building AI systems for recruitment given its spike in employee demand.

But in 2015, they realised the system was processing candidate applications in a discriminatory way. This happened because the data which the models were extrapolating from were resumés submitted over a ten year period, which predominantly came from men; a reflection of institutionalised ine-

quality across the tech industry.

The system deemed male candidates to be preferable and actively suppressed resumés which included the word 'women's' - for example; 'women's hockey team captain'.² Not only this, but it downgraded CVs based on candidates having graduated from all women colleges, and promoted CVs containing verbs more aligned with 'masculine' language such as 'executed' or 'captured', according to inside sources speaking on condition of anonymity. The project has since been disbanded and replaced by another which refocused on diversity. But the question arises; what about all the women who were systematically overlooked between 2014 and 2015? They will most likely never know, which leads onto the next issue in biased AI recruitment: justice. 🇬🇧

AI: the perfect scapegoat

Activists cognisant of the dangers of sexism becoming an underlying basis for hiring are increasingly concerned about transparency in the use of AI.

The Equality Act 2010 makes it unlawful for an employer to discriminate against employees because of their sex, which is what a plaintiff would normally cite as grounds for suing for gender discrimination in the hiring process. However, a prospective employee having their resume discarded by an AI sys-

tem might never know it was being used. Currently, the Computer Fraud and Abuse Act (CFAA) in the US allows the criminal prosecution of bias detectives, researchers and journalists who test websites' hiring algorithms for gender discrimination by uploading dummy CVs for example, which is being challenged by the American Civil Liberties Union. "We are increasingly focusing on algorithmic fairness as an issue," said Rachel Goodman, a staff attorney with the Racial Justice Program at the ACLU.³

Transparency in AI usage is of paramount importance here.

But it's not just hiring which reflects this bias. Facial recognition software embedded in most smart phones works best for those who are white and male, and scoring systems, fueled by potentially biased algorithms, are increasingly being used to make decisions about people's lives in relation to finance, jobs and insurance.⁴ 🚩

Man-guage and AI

A common denominator causally linked to the emergence of embedded bias in AI is language.

This can be from the inclusion of 'man' as the primary prefix or suffix for words, e.g. 'hu-man/kind', 'man-power' (incidentally, the name of the second largest recruitment company), 'man-made', 'workmen', 'man of the house', 'salesman', and the most obvious examples: 'fe-male' and 'wo-men' as being versions of the provided default; aka men. On the other side of the gender spectrum, occupational terms used in relation to women are often pre-modified

by a gender specification such as 'female lawyer' and 'woman judge', identifying their existence as counter to societal expectations. The word 'girl' is frequently used in disparaging and sexualised contexts, for example; 'scream like a girl', the mockery of boys who are 'beaten by a girl' in games, and 'fight like a girl'; phrases which the sports campaign 'This Girl Can' challenged across the globe by re-owning 'girl' and showing women performing powerfully to catalyse a change in paradigm. Joanna Bryson, a researcher at the University of Bath, studied a program designed to learn relationships between words.

It trained on millions of pages of text from the internet and began clustering female names and pronouns with jobs such as "receptionist" and "nurse". Bryson says she was astonished by how closely the results mirrored the real-world gender breakdown of those jobs in US government data, a nearly 90% correlation⁵.

This is problematic in language enough which manifests as unconscious and conscious bias in people, without having the machines we are becoming increasingly dependent on become patriarchal as a result. 🚩

Cyber gender parity: next steps

With regards to AI's perpetuation of sexist gender ideologies via language, the training data for machine learning algorithms needs to be reevaluated and reformed.

Not only this, but legislation needs to be drawn up relating to a standard for this scrutiny and AI organisations legally held to it before data can be utilised.

In terms of policy, government services and companies must also

disclose if a decision has been entirely outsourced to a computer, and, if so, that decision needs to have a definitive legal route for being challenged. Sandra Wachter, a law scholar at the Alan Turing Institute says that the existing laws don't superimpose accurately onto the way technology has advanced. There are a variety of loopholes that could allow the undisclosed use of algorithms. She has called for a "right to explanation"⁶ as in-

cluded in GDPR, which would require a full disclosure as well as a higher degree of transparency for any use of these programs. To enforce this, an auditor for AI use cases needs to be established which can routinely examine AI programs which make critical decisions and actively check for bias, rather than it being discovered accidentally by programmers and news of it quashed by companies concerned about their optics. 🚩

Changing the de facto default design

The one solution which ties all issues of ingrained sexism

within AI is representation. Caroline Criado Perez's seminal work

*Invisible Women: Data Bias in a World Designed for Men,*⁷ is par-

ticularly salient here.

Its findings in women killed and harmed as a result of accidental design that focused on men as the default was staggering and prompted a rightful media storm on the research. A case study of hers wherein the most popular cooking stoves in India were exposing women to the equivalent of four packs of cigarettes worth of smoke a day showed a clear conclusion; as soon as women were included in the design process, better design happened. Researchers created a device made of recycled metal which has been adopted in India and now many more countries. This generated more revenue and resulted in hundreds of thousands of happier, healthier people.

Employers in the tech sector need to incorporate affirmative action

strategies for more female recruitment. Female policymakers, software engineers, developers and users of any AI products need to be present at all stages; proposals, investing, testing, scrutinising, launching and especially as decision makers on company boards. Feminist women are overwhelmingly the authors of AI discrimination research (yours truly; case in point!) and are able to provide the thought leadership and expertise to mitigate the corruption of otherwise brilliant systems. In other words, designers need to let women even the algorithmic playing field.

As an example of this in both the design process and deployment, F'xa⁸ (find it at f-xa.co) is a voice assistant created by NGO The Feminist Internet which teaches users

about AI bias. It also challenges the cloying feminine obsequiousness and servility programmed into virtual assistants like Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's Google Home. Despite the underrepresentation of women in AI development, voice assistants, which are mostly used for aid in mundane tasks are almost always female by default and given feminine names and voices. Researchers also found that Siri responded to 'you're a b*tch' with 'Hey, I'd blush if I could', (I mean, Siri-ously?) which became the title of the UNESCO report⁹ on reinforcing maladaptive gender stereotypes. Apple has since changed this response, but to a neutral one; 'I don't know what you mean' rather than one which asserts that it is inappropriate and unacceptable. 🚩

Policy recommendations for eradicating AI gender bias

- Companies to adopt affirmative action strategies with the aim of gender parity in all stages of AI production.
- The UK to restrict solely automated decision making and profiling which excludes any human involvement (post Brexit) as existent in EU data protection law.
- Full transparency on outsourcing to AI programs in hiring and other decisions which directly impact citizen's lives, such as those based on insurance and household income data. To quote the UK Government AI Select Committee Report: UK: Ready, willing and able?; 'We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take. In cases such as deep neural networks, where it is not yet possible to generate thorough explanations for the decisions that are made, this may mean delaying their deployment until alternative solutions are found'¹⁰.
- An independent regulator to audit AI products under an agreed AI ethics framework which substantially affects individual's lives.
- Consideration of an approval-based regime to be enforced by a regulator to be applied to certain products or sectors which greatly affect individuals' lives.
- The Right To Explanation for those unsuccessful in being hired wanting to challenge the use of AI in the process enshrined in law.
- Diversity quotas enforced for those charged with constructing training datasets Government to make information about the AI systems they use accessible to the public. 🚩

THE RACISM FEEDBACK LOOP: HOW BIG DATA CAN AUGMENT INSTITUTIONALISED PREJUDICE

Another rather insidious manifestation of social inequality has been racial profiling in AI. The data these programmes rely on; arrest records, postcodes, social affiliations, income – can reflect, and further ingrain, human prej-

udice, amplifying the inequalities of our past and affecting the most vulnerable members of our society.

COMPAS (Correctional Offender Management Profiling for Alterna-

tive Sanctions) is a tool used by U.S. courts to assess the probability of a defendant becoming a recidivist (repeat offender), was found by Julia Angwin's investigation to be labelling black people twice as much as white people as recidi-

vists whilst mistakenly inverting the decision for whites. It labelled them low risk despite it being found that they were more likely to reoffend.¹¹ This significantly impacts the decisions of judges to set parole, which translates to a very real infringement on what would otherwise be reasonable releases for people unlikely to recidivate.

“If you’re not careful, you risk automating the exact same biases these programs are supposed to eliminate,” said Kristian Lum¹², lead statistician at the NGO Human Rights Data Analysis Group (HRDAG). Lum

analysed PredPol, which is a program that predicts hotspots where future crime is most likely to occur, and found that it could get stuck in a feedback loop of over-policing neighbourhoods whose residents were predominantly black or brown. She fed the same program Oakland’s drug crime data and it yielded worrying results. The program suggested black neighbourhoods twice as much as white ones, and yet when the cities’ overall drug use was modelled based on national statistics, the hotspots were far more evenly dis-

tributed. Even more worryingly, in simulations depicting what would happen if police had acted on the PredPol’s analysis and increased arrests accordingly, the program would enter a feedback loop, predicting increasingly more crime in the neighbourhoods it told police to visit most, meaning more and more police would be sent in. The racial profiling became a hypothetical self-fulfilling prophecy, as can happen in real life with stop and frisk authorities exploited to racially target ethnic minorities. 🚩

Racism in, racism out - Hamid Khan

The proverb ‘garbage in, garbage out’ is particularly applicable here. If you give programs exaggerated or flawed information with no basis for fixing it, they’ll just process the information and regurgitate it. But rooting out bias can be difficult; more subtle and nuanced than most would expect.

As LexisNexis UK (computer assisted legal research corporation) said, “biases may originate in the data used to train the system, in data that the system processes during its period of operation, or in the person or organisation that created it.”¹³ Research on a machine learning tool named ‘word embedding’ as published in the journal Science found that the AI system was more likely to associate European American names with pleasant words such as “gift” or “happy”, while African American names were more commonly associated with unpleasant words. This bias has also reared its head within Google visual identification algorithms, which could not

distinguish between gorillas and black people¹⁴, and three years after the problem was identified, Google was still unable to fix it. Instead, they simply disabled the ability to search for gorillas in products such as Google Photos which use the feature. Further to this, in 2017 Google’s image recognition was found to be unable to classify Chinese faces properly. Chinese customers were able to open each others’ phones, leading to serious breaches of privacy.

Similarly, when Microsoft launched their AI chatbot ‘Tay’ on Twitter, they underestimated this exact phenomenon, culminating in a scandal and Tay having to be removed from the internet. Within 24 hours, the conversational chatbot had spewed out a whole host of racist and misogynistic tweets, including stating ‘Hitler was right I hate the Jews’ and ‘feminists need to die and burn in hell’¹⁵ thanks to the more depraved interactions from the Twitter and 4chan sewer. Microsoft subsequently launched

Zo, a chatbot preprogrammed to terminate conversations mentioning controversial topics, but the lesson was clear for Microsoft and indeed applicable across AI systems; our world is value-laden, and designers need to be mindful of what values they want their systems to reflect. Humans are often trusted to make these trade-offs between competing values without having to explicitly state how much weight they have put on different considerations. Algorithms are different. They are programmed to make trade-offs according to unambiguous rules. This presents new challenges.

This is not to say that AI does not have potential in this field. If designed with integrity and rigorously and specifically tested for the manifestation of bias, it can and already has certainly improved the speed, quality and neutrality of decision making. 🚩

An issue as its own solution: the data gap

Unusually, the core of the issue is also the core of the solution in this category: data.

The foundational data in these programs can be biased, and therefore corrupt output. It is also typically

unlawful to utilise data pertaining to protected characteristics such as gender, race, etc. as this can be easily used to discriminate. However, in order to evaluate diversity, data on diversity must be collated. ‘This tension between the need to

create algorithms which are blind to protected characteristics, while also checking for bias against those same characteristics, creates a challenge for organisations seeking to use data responsibly’¹⁶.

There is also a significant data gap due to the higher likelihood of white, middle class users owning more devices and having greater access to internet connecti-

ty, making this demographic the epicentre of data gathering. The question of governance also arises here; who should be responsible for governing, auditing and

assuring these algorithmic decision-making systems? 🚩

Policy recommendations: multicultural AI

- Guidance to be established to make AI systems intelligible in terms of accountability, traceability and explainability as well as decision making. AI systems that are outwardly inscrutable to have explanation systems added.
- Increase governmental focus on creating open, diverse datasets, for example identifying gaps in the Open Data Initiative, which are representative of the entire population to close the data gap caused by privately held datasets.
- Datasets to be subject to bias scrutiny at each stage:; planning, training, and deployment.
- A fund to be established by the government for diversity testing, especially in training datasets.
- Government co-operation with data monopolies such as Apple, Google to share issues in datasets and counter-act data dominance.
- Transparency and clear routes for accountability to be required where AI is used to make decisions with significant impact, especially in policing and parole frameworks. 🚩

Moulding AI to champion equality

Whether racism, misogyny or other forms of inequality surfacing in AI, the lessons and recommendations in rectifying them remain largely parallel.

Improving representation is one of the most efficacious solutions in all developmental stages for both

women and ethnic minorities, as well as diversifying datasets for the latter. Transparency is also key, and as AI takes more and more of an adjudicating role in our lives, accountability and ethical standards are imperative for the prevention of societal regression. AI has in-

credible potential to be a champion of equality and ameliorate, rather than compound our imparities. We are at a pivotal point, and we have a crucial window of opportunity to navigate this proverbial ship towards reformation, which we must not disregard. 🚩

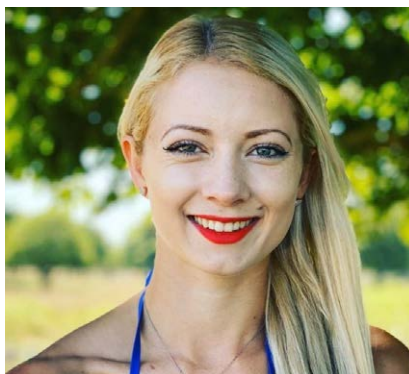


References

1. Entelo, Recruiting Automation Trends Report, 2019, <https://resources.entelo.com/2019-recruiting-automation-trends-report>
2. Aviva Stahl, October 12, 2018 Amazon's now-defunct AI hiring tool was anti-woman <https://womensmediacenter.com/news-features/amazons-now-defunct-ai-hiring-tool-was-anti-woman>
3. Rachel Goodman, Amazon scraps secret AI recruiting tool that showed bias against women, Jeffrey Dastin 2018 <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
4. Nicol Turner Lee, Paul Resnick, and Genie Barton Report: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms Wednesday, May 22, 2019
5. Joanna Bryson, Rise Of The Racist Robots – How AI Is Learning All Our Worst Impulses, 8 August 2017 <<https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses>>
6. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99.
7. Caroline Criado Perez, Invisible Women: Data Bias in a World Designed for Men, 7 Mar. 2019, Harry N Abrams
8. Cara Curtis, This feminist chatbot challenges AI bias in voice assistants, F-xa.co, June 2019
9. UNESCO 'I'd Blush If I Could : Closing gender divides in digital skills through education. <https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1,2019>
10. Select Committee on Artificial Intelligence Report of Session 2017–19 HL Paper 100 AI in the UK: ready, willing and able?, <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> 2019, Government publications
11. Julia Angwin, Jeff Larson, Lauren Kirchner, Machine Bias, 23 May 2016, Pro Publica
12. Kristian Lum, Human Rights Data Analysis Group, Rise Of The Racist Robots – How AI Is Learning All Our Worst Impulses, 8 August 2017
13. Dr Michael Rovatsos, Dr Brent Mittelstadt, Dr Ansgar Koene, Landscape Report: Bias in algorithmic decision making, Centre for Data Ethics and Innovation, <https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making,2019>, House of Lords Publications
14. Amy Tennery, Gina Chereus, Microsoft's AI Twitter bot goes dark after racist, sexist tweets, <https://www.reuters.com/article/us-microsoft-twitter-bot-idUSKCN0WQ2LA2016>, Reuters
15. Hannah Devlin, AI programs exhibit racial and gender biases, research reveals, <https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals>, 13 April 2017
16. Jonathan Vanian, Unmasking AI's Bias Problem, <https://fortune.com/longform/ai-bias-problem/>, June 25, 2018

Bibliography

- [1] Meijer, A. and Wessels, M. "Predictive policing: Review of benefits and drawbacks." *International Journal of Public Administration* 42.12 (2019): 1031-1039.
- [2] Ratcliffe, J. H. "The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction." *Police practice and research* 5.1(2004): 5-23.
- [3] Couchman, H. "Policing by Machine", *Liberty* (2019) <<https://www.libertyhumanrights.org.uk/issue/policing-by-machine/>>
- [4] Perry, W. L. et al. "Predictive Policing", *The RAND Corporation, Safety and Justice Program* (2013)
- [5] Babuta, A. and Oswald, M. "Data analytics and algorithmic bias in policing", *Briefing Paper, Royal United Services Institute for Defense and Security Studies* (2019) <https://rusi.org/sites/default/files/20190916_data_analytics_and_algorithmic_bias_in_policing_web.pdf>
- [6] Kutnowski, M., "The Ethical Dangers and Merits of Predictive Policing", *Social Innovation Narratives, Journal of CSWB, VOLUME 2, NUMBER 1* (2017) <<https://journalcswb.ca/index.php/cswb/article/view/36/75>>
- [7] "Ethics Advisory Report for West Midlands Police", *The Alan Turing Institute* (2017) <https://www.turing.ac.uk/sites/default/files/2018-11/turing_idepp_ethics_advisory_report_to_wmp.pdf>
- [8] Mozur, P., Zhong R. and Krolik, A., "In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags, *New York Times* (2020)
- [9] Meeting notes, Thirty-second SAGE meeting on COVID-19 (2020) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/888807/S0402_Thirty-second_SAGE_meeting_on_Covid-19_.pdf>
- [10] Radaelli, L. et al., "Quantifying Surveillance in the Networked Age: Node-based Intrusions and Group Privacy", *arXiv preprint* (2018)
- [11] Stokel-Walker, C., "Coronavirus has ushered in the terrifying era of coughing assaults", 2020 <<https://www.wired.co.uk/article/coronavirus-coughing-spitting-assaults>>
- [12] Abbas, R. and Michael, K., "The coronavirus contact tracing app won't log your location, but it will reveal who you hang out with", *The Conversation* (2020) <<https://theconversation.com/the-coronavirus-contact-tracing-app-wont-log-your-location-but-it-will-reveal-who-you-hang-out-with-136387>>
- [13] Kerr, B., "One in five shares data from Smittestopp", *Norway Today* (2020) <<https://norwaytoday.info/news/one-in-five-shares-data-from-smittestopp/>>
- [14] Findlay, S., Palma, S. and Milne, R., "Coronavirus contact-tracing apps struggle to make an impact", *Financial Times* (2020) <<https://www.ft.com/content/21e438a6-32f2-43b9-b843-61b819a427aa>>
- [15] CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment (2019)



Cecilia is a Policy Manager for HMRC. Her piece was shortlisted for the Outstanding Contribution to Tech Regulation Award by Cog X; a global tech leadership summit.

Twitter: @CeciliaEve4

ARTIFICIAL INTELLIGENCE IN HEALTHCARE: BRITAIN 2020

By Mohamed Hameda

Medical data is being produced at such a rate where it is projected to double every 73 days by 2020¹, and with the ability for an individual to create the equivalent of 300 million books of health-related data in their lifetime² and the vast improvement in AI performance within healthcare, the opportunities of AI to enhance and add value to our landmark healthcare system and firmly bring it into the 21st century are endless.

The already data-rich medical field combined with other data (such as air pollution, humidity, air pressure) alongside the latest AI techniques can discover hidden trends we never knew existed and find out more about our biology like we have never been able to do before. However, with such an opportunity comes ethical challenges and questions around the potential risks. This part of the pamphlet will

outline such risks as well as steps for the future to ensure that Britain leads the way in ethically using AI to improve the healthcare and lives of everyone.

The conversation is well underway on how AI can be used within the NHS, with the Secretary of State for Health and Social Care unveiling plans in 2019 to launch a National Artificial Intelligence Lab for the NHS as part of a £250m investment³ and a further announcement of a £140m competition to speed up the delivery of AI technologies which can be rolled out within the NHS⁴. Whilst this investment is ambitious and welcomed, we must also note that trust is a crucial part in ensuring we maximize the potential of such investments. The NHS is a national institution, and if we are to transform it in the way we want to see it transform, it can only be done with the confidence of the public.

What must also be kept in mind about investment of AI within healthcare is that it isn't just an investment in providing new ways of delivering healthcare, but it is also an investment in the quality of current provisions and can also provide significant savings. Investment in AI does not have to be an investment to replace what we already have, but it can provide another tool to healthcare professionals: giving practitioners a "second pair of eyes"; giving doctors proportionally massive head starts on life-threatening illnesses; and the ability to conduct research to find out relationships between behaviour, diet, demographics and many other factors in relation to an individual's health that would provide critical medical intel which wouldn't exist otherwise. If we allow humans to do what they do best and allow machines to do what they do best, we will maximize the potential of AI within health. 🇬🇧

Moving from reactive to preventative healthcare through the use of data

Our healthcare system is a reactive one, as are those in many other parts of the world. We encourage a healthy lifestyle through regular exercise, a nutritious diet, and discouraging bad habits. But, these guidelines, whilst preventative in nature, are only to reduce the general risk to ill-health rather than identifying the specific risks of certain illnesses and associated targeted care.

When an individual is ill, they may not actually know they are ill, as symptoms may not show or are not strong enough to be noticeable. This is where the "clock" starts for doctors. Once symptoms become identifiable, the individual may wish to see a professional,

but waiting to see a professional is time that professionals are losing to treat a patient. Once a patient is seen, further time is lost due to the wait for tests and results. Once the patient is finally diagnosed with an illness, it is sometimes too late and such overheads, whilst necessary, have taken too long and caused a loss of effectiveness in potential healthcare. This workflow is clearly not ideal, particularly with additional pressures that are to come onto the NHS in the coming years. With the use of AI, such pressure can be suppressed and efficiencies gained.

A common theme amongst emerging AI in healthcare is the ability to detect sickness before a profes-

sional can identify and diagnose an illness. One case in point⁵ is a piece of software developed in 2016 by researchers at Houston Methodist, which is a form of AI that interprets mammograms, assisting doctors by providing predictions of breast cancer risk. It translates patient charts into diagnostic information at 30 times the speed of a human with 99 percent accuracy. Results like this reflect several benefits that occur from similar AI technology, including but not limited to:

- Relinquishing the need for further diagnostic tests: in this case study, reducing the number of unnecessary breast biopsies
- Reducing the number of false

positives and false negatives: in this case study, this is due to physicians being given more information to better assess the cancer risk and thereby if a biopsy is required, reducing the number of unnecessary breast biopsies

- Quicker diagnostics, due to the freeing-up of occupied slots for further tests whilst also saving costs (in this case study, slots for breast biopsies)
- More physicians' time made available by reducing unnecessary tests. This relieves work pressure and allows them to

spend more time on patients who are more at risk, improving the overall quality of patient care

- The reduction of unnecessary tests also reduces unnecessary anxiety, improving a patient's mental health and instilling confidence in the healthcare system

All these efficiencies make better use of staff time, provide better quality care to patients, and enable the reallocation of NHS resources to other departments.

It is estimated that late diagnosis of

four common cancers (colon, rectal, lung and ovarian) costs the UK £165 million a year⁶ (brought down to £111 million, if the number of cancers diagnosed at a late stage were halved) and this doesn't even take into account the hidden costs such as those associated with a patient and their loved ones' mental health when having to go through the waiting process. Investment in the short-term can provide long-term savings through more effective and efficient output from medical staff and lead to a better quality of healthcare. 🚩

Cancer, coronavirus, and convolutional neural networks

Cancer is a horrible condition which not only heavily occupies healthcare services but can be traumatic for individuals involved and associated with someone who has cancer, especially as it is one of the leading causes of deaths in the UK (in 2017, it was the leading cause of death⁷) by broad disease group for both men and women in every nation of the UK. This problem is compounding, with the projected number of deaths caused by cancer to reach over 13 million in 2030 worldwide⁸ (not taking into account the knock-on impacts of coronavirus).

However, with 30% - 50% of cancers being preventable through early detection⁹ (as well as prevention measures such as healthy lifestyle and public health measures), we can save a significant number of people's lives. Cancer is diagnosed late for a number of reasons, which include delays in obtaining appointments at the hospital and delays in GPs referring patients on for tests and treatment as well as a lack of awareness of the signs and symptoms of cancer¹⁰.

Moreover, this issue of obtaining early diagnosis of patients with cancer, and other conditions, will be amplified by an unprecedented scale by the legacy of coronavi-

rus rippling through the NHS. The British Heart Foundation estimates that since the start of the outbreak, 28,000 procedures were delayed in England by June 2020¹¹, putting a serious number of patients at risk of permanent complications. As Dr Sonya Babu-Narayan, associate medical director at the British Heart Foundation puts it: "People with heart and circulatory diseases are already at increased risk of dying from Covid-19, and their lives should not be put at even greater risk by missing out on treatment for their condition. If hospital investigations and procedures are delayed too long, it can result in preventable permanent long-term complications, such as heart failure".

Furthermore, according to Cancer Research UK¹², as of the 1st of June 2020, 2.4 million people in the UK were waiting for cancer screening, further tests, or cancer treatment. They also estimate that during this time, 3,800 cancers would normally be diagnosed through screening and a further up to 20,300 cancers identified through urgent cancer referrals. Cancer Research UK also state that there will be a build-up of treatment waiting to be carried out, with 12,750 fewer patients receiving surgery, 6,000 fewer for

chemotherapy and 2,800 fewer for receiving radiotherapy since lockdown started (all of which will need to be carried out as soon as possible). These figures illustrate the fears of health bosses that due to the crisis, the number of treatments waiting to be carried out under the NHS treatment could double to 10 million by the end of the year¹³ (assuming that the health service steadily returns to full capacity within a year). With such unprecedented pressure on the NHS, it is going to require an unprecedented relief package. Such a package can include utilising AI technology to realise enormous efficiencies where possible as part of the recovery strategy.

Further promising signs are also coming from the development of AI technology, with a paper being published at the beginning of 2020 showing that an AI algorithm outperformed all of the human readers in reading mammograms, in an independent study of six radiologists¹⁴. In the study, comparing the performance of the algorithm with the UK system of examining mammograms (double reading process; where two radiologists review a mammogram, with a third radiologist reviewing the mammogram if there is a disagreement), the AI

model was just as good in detecting patients with cancer, with the AI reducing false positives by 1.2% and false negatives by 2.7%, reducing the number of patients who are wrongly diagnosed (which reduces unnecessary treatment and anxiety in patients) and reducing the number of sick patients who do not end up receiving treatment respectively. When the AI model was used as part of the double reading process, it reduced the workload of the second reading by 88%. With The Royal College of Radiologists identifying a shortfall of at

least 1,104 radiologists across the country¹⁵ it is critical that we find innovative solutions in response to the shortfall we originally had and in addition to the legacy that will be left behind by the coronavirus crisis. One requirement of the AI model that would need to be fulfilled before deployment would be that clinical trials need to be carried out beforehand, to test their effectiveness in practice, simply because it has not been through the rigour of clinical trials, like many other similar “breakthroughs”. This regulation is an issue that the gov-

ernment must make a decision on, and the current lack of clarity is currently holding back progress.

With the extraordinary pressure that is to come onto the NHS, technology such as the ones discussed could play a massive part during recovery from the coronavirus crisis, improving the health-care of patients whilst freeing up health-care professionals to do other critical tasks, and most of all, allow us to move our healthcare system from being reactive to being proactive. 🇬🇧

The internet of things and healthcare

In addition to AI being used to analyse outputs from medical tests, which show a person's health at a singular point in time, we can also harness the data collected from wearable technology over time.

A common FitBit worn on someone wrist can collect a number of pieces of information¹⁶ such as: “number of steps you take, distance travelled, calories burned, weight, heart rate, sleep stages, active minutes, and location”, and this is just the tip of the iceberg in terms of what can be collected from all types of wearable technology.

With data collected from wearable technology, in combination with test results and other data external to a human body, such as temperature, humidity, and pollution levels, we can discover hidden correlations which we would have never noticed otherwise. Action which is of cause for concern can also follow from data. For example, a dramatically reduced heart rate can be followed up with automated contact with the emergency services and a downturn in mood of a person could be followed with contact with their best friend or close relatives.

This abstract model can be referred to as the “Internet of Things” (IoT) in healthcare, where different

devices and data from different sources are interconnected over the internet. This data can then be analysed, generally in real-time, responding in the interest of an individual's health. Having such information and analysis about an individual is of benefit to both healthcare practitioners and the data subject. Real-time feedback based on more complete personal datasets can give instant benefits to individuals through, for example, instantly displaying how certain actions and behaviours can affect their health and what behaviours and habits should be discouraged or encouraged.

Once those recommendations are then carried out by the individual, the impacts can then be monitored to aid future recommendations as part of the training process of the AI. The results lead to the improvement of the health of both the individual and also of the recommendations given to others. The collection of data doesn't just aid professionals in understanding where a patient's health is, but it can also help in seeing the effects of given treatments to patients, giving professionals a whole lot more information than the current process, which is gathering information from follow up appointments. This is provided through continual monitoring of a patient through-

out their everyday life, to evaluate how effective a given treatment is on a person's symptoms, giving more power to professionals and researchers and a better understanding of the treatment given to a certain individual.

One example of such infrastructure being built is IBM Watson Health, which takes data from different data sources about an individual data subject (e.g. a patient), uses the relationship between different data (about the data subject), and incorporates research from the field of medicine in its algorithms to produce recommendations to health-care professionals. The idea is to use the latest advances in the field of medicine in combination with analytics on a large dataset (with the cognitive ability of IBM Watson) to exist all under one roof, in the cloud. Unfortunately, IBM Watson progress still seems to have some way to go, with progress in the past showing disappointing results, one case concluding with “IBM's Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments”¹⁷.

However, other commentators are more optimistic. A report by Stanford University¹⁸ in 2016, titled “Artificial intelligence and life in 2030”, notes that “emerging (inter) connectedness between the home environment and health-monitor-



ing devices, has created a vibrant new sector of innovation. By combining social and healthcare data, some healthcare apps can perform data mining, learning, and prediction from captured data, though their predictions are relatively rudimentary. The convergence of data and functionality across applications will likely spur new and even obvious products, such as an exercise app that not only proposes a schedule for exercise but also suggests the best time to do it, and provides coaching to stick to that schedule.”

Additionally, the report lays out a vision for elderly care, suggesting technology in three categories:

- Life quality and independence - this is on smart devices in the home, which will help individuals who require aid in daily tasks essential for living, for instance: cooking, dressing, and toileting. Additionally, “predictive analytics” can be used to predict when the best times are to suggest certain actions, such as: “‘nudge’ family groups toward positive behaviors, such as reminders to ‘call home.’”
- Health and wellness - on the theme of encouraging healthy

lifestyles, the report mentions mobile applications that monitor an individual’s activity. With the help of social media, recommendations can be made to such individuals to change certain behaviours to stimulate better mental and physical health. It also mentions health monitoring at home, where carers and close friends and family of an individual can be alerted if there is a change in mood or behaviour of that individual which suggests something is not right.

- Treatments and devices - talks about a range of assistive equipment having added on capabilities to improve an individual’s quality of life: “Physical assistive devices (intelligent walkers, wheel chairs, and exoskeletons) will extend the range of activities of an infirm individual” as well as additional medical equipment to enhance individual treatment and aid in increasing the capacity of the health-care system, for instance: “Personalized rehabilitation and in-home therapy will reduce the need for hospital or care facility stays”.

The ideas above show how the IoT can be used for protecting, en-

couraging, and facilitating good health of individuals and family, which in turn ensures more people live healthier, happier lives and are only receiving health-care when needed. This is all achieved through utilizing data about an individual from a wide range of data sources and bringing it all together to add another layer to our knowledge about an individual’s health. Additionally, the guidance that can be given from these technologies (witness the seemingly limitless fitness-centred products that are currently being sold) and the potential of the new and better ways for individuals to understand their own health can provide countless benefits, including: outlining the consequences of an individual’s actions on their health, both positive and negative; what the current state of their health is, and what actions they can take to improve their health. The continual collection of data from different data sources adds several dimensions to what a professional can see about a patient’s health, especially after prescribing a course of action to a patient. This gives professionals a whole lot more data, deeper insights and allows them to deliver better quality healthcare compared to what can be achieved in a short follow-up appointment. 🚩

Revolutionising emergency care

Individuals, especially the elderly and those who live alone, can now use wearable technology to send an alert to their neighbour, family members or even emergency services, such that steps can be taken to protect an individual who might be vulnerable in case of emergency.

Additionally, the technology may not need to actually wait for the incident to actually occur, but could trigger an alert as soon as there is a suspicion or high likelihood that an imminent event is about to occur (such as a heart attack), which is life-changing in terms of caring for the elderly and emergency re-

sponse. Predictive recommendations are one of the key benefits of modern AI technology. 🚩

Case study - Acute Kidney Injury and Deepmind

Acute Kidney Injury (AKI), without early detection, can often be lethal. AKI is where the kidney abruptly ceases to function correctly, the consequences of which can lead to permanent kidney damage and death¹⁹. AKI affects 1 in 5 people who are admitted to hospital via emergency and around 100,000 lives are lost each year²⁰ with research showing that up to 30% of cases are preventable through earlier recognition²¹.

Google Deepmind has developed an AI model for which they claim “in the future, could give doctors a 48-hour head start in treating acute kidney injury”. The AI model currently “predicts 55.8% of all inpatient episodes of acute kidney injury, and 90.2% of all acute kidney injuries that required subsequent administration of dialysis, with a lead time of up to 48 hours and a ratio of 2 false alerts for every true alert”²². So, the technology developed has an accuracy of 33% (one is every three alerts is a true alert) and sensitivity of 56%. These are promising early numbers, but these results must be approached with caution, because as always, the devil is in the detail. The data that was used to train the AI was from a “multi-site retrospective dataset of 703,782 adult patients from all available sites at the US Department of Veterans Affairs (VA)—the largest integrated health care system in the United States²³”. But, the gender of the individuals whose data was used was almost all men, with the dataset of veterans consisting of 93.6% males²⁴. As described in other parts of this pamphlet, this lack of representation within a dataset can lead to further unintended consequences, including the model performing much worse for specific demographics. Looking into more detail in the proposed model and the claim of giving “doctors a 48-hour head start in treating AKI”, we see that with an accuracy of 33%,

the sensitivity in the range of 42 to 48 hours in the time between prediction and AKI onset is about 10% and we only get a sensitivity of 56% in the range of 0 to 48 hours with the same accuracy²⁵.

These results show having to put up with a lot of false alarms before identifying an impending AKI onset with the accuracy given. With the sensitivity reported, around half of AKI onsets will not be picked up (through false negatives). We must acknowledge that these are preliminary results with signs of potential, however. A peer-review into the model (with Deepmind’s Stream app) suggested that the average cost of admission of a patient with AKI would be lowered by 17%, which would lead to a saving of £2123 per patient (not including the cost of providing the app or the cost of long-term dialysis in patients with AKI that goes untreated²⁶). But, it noted that “e-alerts alone ‘might fail to improve outcomes’ and appropriate training was needed for a digital solution to be successful²⁷”. To really test its effectiveness in practice, clinical trials would need to be conducted to see the added value such technology can bring and verify its effectiveness, with such preliminary results not being sufficient to clinically verify the claims by Deepmind. We can therefore conclude that Deepmind’s AI model is a good start, with a lot of potential, but there is still a lot more refinement that is needed and requires extensive testing to be a credible tool in healthcare. Prof Paul Leeson, Professor of Cardiovascular Medicine at the University of Oxford, commented on the results of the trial carried out by Deepmind stating: “This is important work in which the team have overcome several technical challenges to show it is possible to successfully apply AI to large scale electronic health records. Trials are still needed to test whether this early warning is useful

to doctors to improve patient care, without causing too many false alarms, or missing patients that the AI also overlooked. However, this is another strong example of how AI appears to have the potential to augment delivery of healthcare.²⁸”

This also brings another problem to light, which is actually testing how effective AI technology is in the field of medicine. The issues are due to the serious consequences when it goes wrong for patients, with a false positive leading to treatment that is not necessary, wasting resources (with potentially invasive treatment being carried out on patients unnecessarily) and a false negative leading to someone being mistakenly identified as being healthy, which can lead to serious consequences²⁹. Furthermore, because of lack of regulation, NHS trusts may make direct use of AI technology without such technology being clinically tested. This is summarised by Dr Franz Kiraly, Honorary Lecturer at UCL in the Department for Statistical Science, who stated: “Due to lack of regulation in the area, unlike for pharmaceutical drugs, it will probably proceed to direct use in some hospitals, rather than a clinical study”³⁰.

Dr Franz Kiraly also argues that the techniques and algorithms used by DeepMind (such as deep learning) may not be as suitable as simpler approaches. He states: “The paper provides some evidence that AI algorithms are capable of early warnings, and not that the DeepMind algorithms are necessarily the best choice to do so. Deep learning, or neural network algorithms, have disadvantages including that they are very resource demanding (they need AI experts, hardware, and time to run the algorithm), it is difficult to understand why a recommendation by an algorithm is made and they can be prone to failure outside of the range of data they



were trained upon. The alternatives to deep learning algorithms are simpler approaches, such as systems involving logistic regression models, which are fast, readily available and easily interpretable and well-understood by doctors. From a healthcare perspective, it may be preferable to use a logistic regression model to a deep learning algorithm as long as they were similarly successful in predicting an outcome correctly³¹. Being able to utilise simpler techniques, given they provide similar or better performance, would be advantageous. They are less resource intensive, allow for extrapolation and allow professionals to better understand the technology and how it comes to specific recommendations. Their use improves transparency, ensuring that professionals can evaluate such recommendations and giving them the confidence to challenge and reject recommendations made by AI and therefore ensuring professionals remain in charge (as will be discussed in more detail later).

Being able to predict patient deterioration before any visible symptoms appear and reaching Deepmind's ambition of giving doctors a "48-hour head start in treating acute kidney injury (AKI)" would be a game changer, with the potential of saving hundreds of preventable deaths as well as averting permanent damage to patients' health, whilst providing significant savings to the NHS through early detection.

Saving more lives

Further applications can be found

in elderly care. One pilot program in 2010 by Keystone³² at TigerPlace Independent Living in Columbia, Missouri found that implementing a monitoring-based AI platform which learns patient behaviour and forewarns them of health risks showed an average length of stay being 1.72 years longer for elderly individuals living with the technology.

We also see advances in the area of orthopaedics. For instance, the US FDA (Food and Drug Administration) authorised a piece of technology in 2018 developed by Imagen's, called OsteoDetect, which is a tool that can quickly detect distal radius wrist fractures³³. The FDA are ambitious about the promotion of AI in healthcare, with Scott Gottlieb, former commissioner of the FDA, stating in 2018 that: "Eventually, AI tools could be integrated directly into smartphones or wearable devices for a variety of early detection applications, reducing the need for expensive specialist visits while increasing the likelihood that we're catching potentially serious problems early,"... "These are no longer far-fetched ideas."³⁴

The innovations and technology that has been described is just the tip of the iceberg, with many more ideas, discoveries and announcements being publicly announced on a regular basis. Whilst we must take every reported "breakthrough" with a pinch of salt and look carefully as to how it would fit in our healthcare system and wider society, the opportunity to strengthen our NHS is there for the

taking. What we are seeing is very promising and the potential for the future is unimaginable, in a good way.

A vast majority of the technologies outlined above have a common theme. They not only reveal a better understanding of and provide insights into a patient's health, but give recommendations and suggestions to avert preventable diseases. They don't just respond to symptoms, they stop them from happening in the first place. Such technologies provide a whole new dimension in the fight against preventable diseases, moving our healthcare system and public health messaging from being reactive to preventative through the use of data. Investment in AI solutions can improve health and social care, relieve pressure on NHS staff, and bring serious amounts of savings to the NHS - a healthcare system which the National Audit Office concludes isn't sustainable and "is treating more patients but has not yet achieved the fundamental transformation in services and finance regime needed to meet rising demand"³⁵. With people living longer, demographic pressures putting additional strain on the NHS, the Government currently underfunding services to keep provision at current levels of performance in the future³⁶, and with the legacy of coronavirus on the economy and the NHS, innovative solutions are going to be needed across all areas of government. A big part of that can be in health. 🇬🇧

CHALLENGES POSED BY SUCH OPPORTUNITIES

Building public confidence in AI

Trust is paramount to quality healthcare, and there is evidence to suggest that from a clinical perspective, better quality of life is achieved when patients have higher trust in their healthcare professionals³⁷. However, the public overwhelmingly have reservations on who they trust to conduct analysis and research using their health-related data, as a research carried out by YouGov of over 2,000 UK adults shows³⁸:

- 13% of the public trust multinational tech companies to handle sensitive health data in a confidential manner
- One in ten of respondents (11%) said they are happy for NHS data to be analysed by businesses that do not pay tax in the UK
- 69% raised concerns about this information being analysed in other countries with different laws governing data security and confidentiality
- 76% of the public say that the UK needs a strong domestic AI sector so that it does not have to be outsourced internationally

The research by YouGov also showed that “86% of people said the NHS should benefit from data analysis, with 81% saying explicitly that the Government should act to ensure that the NHS and taxpayers benefit financially”. From these

findings, there is public enthusiasm to use data analysis to improve healthcare (which would coincide with AI technologies), but how it is done is of great concern to the public and therefore shouldn't be taken lightly. Additionally, healthcare staff also have their reservations, with a YouGov poll of 1,027 healthcare professionals³⁹ showing just 12% of NHS staff and private healthcare workers being comfortable with a multinational company carrying out analysis on patient data and only 17% would trust multinational big tech companies to handle such data in a confidential manner. What we can further note from the research is that healthcare workers are as optimistic as the public with 81% supporting the analysis of anonymised data to allow for quicker diagnosis and more effective treatment for patients.

Research in America shows similar patterns, with just 20% of consumers trusting AI-generated advice for healthcare⁴⁰.

From these results, we can conclude that there is public enthusiasm for the NHS to take on new technology (as well as among NHS staff) and make use of data analytics to provide better health-care, with three main areas of concern to the public as NHS staff:

- How ethically do companies that may provide solutions to the NHS behave? For example,

are these companies paying UK taxes and is patient data fully anonymised and handled in a confidential manner?

- Is the domestic AI sector within the UK strong enough - which carries lower risks than outsourcing?
- What data security rules govern patient data (especially relevant, with GDPR no longer being part of UK law as the UK leaves the European Union)?

The latter has been in the public attention with incidents including the Information Commissioner's Office conclusion in 2017 that a “Google DeepMind trial failed to comply with data protection law”⁴¹ and with privacy concerns surrounding the proposed NHS track and trace app raised by politicians (including Harriet Harman MP, Chair of the Joint Committee on Human Rights (JCHR), which “has drafted a Bill which will lay down the purposes for which the data can be used and prohibit its use for anything else”⁴²) (see *Section 1: Law, Order, and Governance*).

What must follow is action to reassure the public and to build confidence among patients and professionals. That is achieved by implementing responsible policy to ensure that Britain leads the way in finding an ethical route of implementing AI within healthcare. 🇬🇧

Ensuring professionals remain in charge and the importance of “the human touch”

Current mainstream opinion is that we must not allow AI solutions to dictate our behaviour, where professionals become too reliant on AI solutions and we head towards the situation where AI solutions become the decision maker in healthcare. Rather, they should be confirming or challeng-

ing any decision made by a professional as a “second opinion” that professionals can utilise (the advantage being that these are a scarce resource for professionals and patients) to critique and evaluate decisions that they make and reverse decisions which are incorrect.

The current generation of AI models do not (yet) have natural intelligence like humans, but either follow sophisticated algorithms using the data provided to them or operate in a relatively obscure and uninterpretable way. The downside of not handing decision making to such machines is that for the short

term, we would be still reliant on the workflow of the healthcare system. However, the long term benefits mean we can adapt better to any inaccuracies and mistakes by both AI solutions and healthcare professionals and learn from a gradual implementation process.

The author's personal opinion is that AI far from matches human intelligence, in that it is lacking in other aspects such as self-awareness, human emotion, social skills, and the ability to handle changing aspects of a problem. We additionally run the risk of anchoring bias, a cognitive bias, where humans become too dependent on an initial piece of information received, thereby becoming unconscious to the fact we become more reliant on the output of AI. We need to remember that the way we unlock the true potential of AI within healthcare is if we allow humans to do what they do best and allow machines to do what they do best.

AI has its limitations in which it performs sub-par to humans in certain aspects of delivery of healthcare, a case in point made by the Royal College of General Practitioners⁴³, which states that "GPs are highly-trained medical professionals: [apps and algorithms and GPs] can't be compared and the former may support but will never replace the latter". Healthcare professionals have not only spent a number of years attaining knowledge and the ability to retrieve that knowledge for the benefit of patients, but they've also learnt a whole host of other skills, such as dealing with patients when delivering

traumatic news. Examples include: telling someone that a loved one has passed away, informing a patient they have tested positive for HIV, or delivering good news, such as when cancer treatment is no longer needed. Making judgments when providing healthcare to a patient with complex needs is an incredibly tricky task. It requires empathy, the ability to communicate and the ability to reconfigure a solution, which current AI models simply don't have. Where a patient works in a job that is a barrier to treatment, has a disability or even isn't fully disclosing information to a healthcare professional out of fear, discomfort or embarrassment, a human might be able to pick up and deal with such scenarios, whereas a AI model can't. Moreover, if a patient is actually providing incorrect information as an input to an AI model, without the AI model recognising the information provided is incorrect, that could be detrimental to an individual's health and could even be lethal, either through wrong treatment or a missed diagnosis. This is exactly why we must remember that humans have human skill which is trained and obtained that no current AI solution can have. AI solutions can support but shouldn't (yet) replace trained healthcare staff.

Furthermore, one particular issue with healthcare given to patients is the ability for a patient to recall key information. For instance, a study out of Brown University School of Public Health⁴⁴ presents a study in which: "Overall, 49% of resolutions were recalled freely and

accurately, and an additional 36% were recalled accurately with a prompt. Fifteen percent could not be recalled or were recalled erroneously. The numbers were similar when medical and behavioural resolutions were examined separately." As previously mentioned, this can heavily skew the results that AI technology gives and even bring heavy differences in the quality of healthcare that is provided to an individual depending on age, anxiety level, and reluctance to disclose symptoms. This is where the ability of a healthcare professional is especially important - in bringing comfort and providing a safe space to a patient for them to disclose symptoms.

However, in all cases specified, data analytics can still be used by healthcare professionals to better inform them - combining the power of data analytics with the wealth of experience of the professional, ensuring we obtain the best of both worlds. To enable this, we must ensure that professionals feel confident to challenge the output of a piece of technology with the knowledge of how it comes to a particular decision (raising difficulties in the case of "black boxes") and ensuring they have a basic understanding of the underlying way these pieces of technology works so as to learn their shortcomings. The reason for such caution is simply in the interest of patient safety. Ensuring professionals stay in charge, we obtain the best of both worlds and no one is left behind, whatever the state of their health.



Quality of data, privacy issues and quashing bias

So far, we have discussed the potential for AI and how stakeholders feel about its use, but we have not actually talked about its own limitations. The three that present themselves the most to the author are: the ability to obtain quality data in high volumes; the subsequent privacy issues that

arise; and the bias that can come with these technologies (the latter two being covered in detail in Section 3: Conclusions: Policy Proposals and Section 1: Equality and Biases).

On the first issue of data quality: for AI solutions to work as intend-

ed, they are largely dependent on large levels of high-quality data. When describing data as being of "high quality", there is no formal definition, but it is generally accepted to be reliable, complete, and consistent. Unfortunately, the data available in the NHS is disconnected across many different

systems. The perfect setup is each patient having a single electronic health record, consistent with every other patient's electronic health record, all under one framework - but, this is simply not the reality in the NHS. A report by Ernst & Young⁴⁵ states: "there will be a significant process and technology costs associated with aggregation, cleaning, curating, hosting, analysing and protecting the transformation of these raw data records into a consolidated longitudinal patient-level data set". However, the report same also notes that the data held by the NHS is of high value, noting that: "The curated NHS data set is an intangible asset with a current valuation of several billion pounds and a realisation of £9.6bn per annum in benefits (i.e., the NHS benefits worth £5bn per annum and the patient benefits worth £4.6bn per annum) that could be unlocked following the generation of insights"⁴⁶. It further estimates that the value of emergency medical record data per patient is greater than £100⁴⁷. This alone shows that the data held by the NHS is a national asset to be recognised and protected and patients are well within their rights to expect maximum returns from their data. However, such value coming from NHS data raises a further issue - that of privacy concerns.

As the need for high volumes of "quality data" increases (companies as well as the NHS are increasingly "data-needy" from wanting to quickly develop working AI solutions), the risk of overlooking privacy concerns of the public increases and can inflate a previous challenge mentioned earlier, which is building public confidence. A case in point is shown in an article in *The Times*⁴⁸ at the end of 2019, where it was reported that Amazon was given free access to NHS data and that "The \$863bn company (Amazon) can access 'all healthcare information' gathered by the NHS at the UK taxpayers' expense, including 'symptoms, causes and

definitions'... Amazon can use the information to make, advertise and sell 'new products, applications, cloud-based services and/or distributed software' and can share it with third parties". The *Times* also stated that "A commercial lawyer who analysed the contract said - 'the most alarming thing is that Amazon isn't paying anything for this and the data is very valuable'". When the deal between the NHS and Amazon was first struck around July 2019, it raised many public concerns⁴⁹. Phil Booth, co-ordinator at medConfidential said: "personal and health data is heavily protected under GDPR in the UK, but Amazon Alexa doesn't comply with the same laws", with Big Brother Watch "labelling the partnership a 'data protection disaster waiting to happen'". This isn't an isolated incident - as mentioned earlier, there are conclusions from the ICO that the "Google DeepMind NHS app test broke UK privacy law"⁵⁰ and news stories showing NHS trusts giving Google the ability to process confidential patient data⁵¹ in combination with Google declining to release contract details. This lack of transparency will ultimately break public confidence and lead to a evaporation in trust between patients and the NHS, which would have several follow-on consequences, as well as killing the potential discussed up until this point.

Lastly, the issue of bias is introduced. As AI solutions are only as good as the data fed into them, we see countless examples of algorithms displaying existing inequalities and even exacerbating such inequalities due to the data being used to train such AI solutions incorporating existing biases in the world today. One example of this having real life consequences is in America⁵², where an algorithm used in many US hospitals was found "less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients

with complex medical needs." - introducing a two-tier system in healthcare and showing serious risk of perpetuating inequality to a level seen when racial segregation was prevalent in the United States. The data that is being used to train such algorithms must be closely examined and scrutinised before being used and even once a result is generated such results must be further dissected. The challenges around data privacy are discussed in Section 3: Conclusions: Policy Proposals.

Not only are individuals' privacy at risk, but there is also a real risk of unmonitored AI solutions taking existing institutional racism within society and showing, perpetuating, and even expanding it. This means that safeguards must be put in place. Opponents to safeguards might call it "red tape", "unnecessary bureaucracy" and say it slows down the implementation of AI solutions, but such measures being put in place will mean the NHS will make better use AI solutions quicker, as where public confidence supports it and individuals feel safe with its implementation the NHS will ultimately get much more support in the long-run.

We must keep in mind that it is the general public and NHS staff who will be working with AI solutions and who are the biggest stakeholders in all this. We need them to play a central role in the implementation of AI solutions in all areas of life, as well as health, if we are to fully realise the potential of AI within existing society.

So far, we have discussed the potential for AI and how stakeholders feel about its use, but we have not actually talked about its own limitations. The three that present themselves the most to the author are: the ability to obtain quality data in high volumes; the subsequent privacy issues that arise; and the bias that can come with these technologies (the latter two being covered in detail in Section 3: Conclusions:

Policy Proposals and Section 1: Equality and Biases).

On the first issue of data quality: for AI solutions to work as intended, they are largely dependent on large levels of high-quality data. When describing data as being of “high quality”, there is no formal definition, but it is generally accepted to be reliable, complete, and consistent. Unfortunately, the data available in the NHS is disconnected across many different systems. The perfect setup is each patient having a single electronic health record, consistent with every other patient’s electronic health record, all under one framework - but, this is simply not the reality in the NHS. A report by Ernst & Young states: “there will be a significant process and technology costs associated with aggregation, cleaning, curating, hosting, analysing and protecting the transformation of these raw data records into a consolidated longitudinal patient-level data set”. However, the report same also notes that the data held by the NHS is of high value, noting that: “The curated NHS data set is an intangible asset with a current valuation of several billion pounds and a realisation of £9.6bn per annum in benefits (i.e., the NHS benefits worth £5bn per annum and the patient benefits worth £4.6bn per annum) that could be unlocked following the generation of insights”. It further estimates that the value of emergency medical record data per patient is greater than £100. This alone shows that the data held by the NHS is a national asset to be recognised and protected and patients are well within their rights to expect maximum returns from their data. However, such value coming from NHS data raises a further issue - that of privacy concerns.

As the need for high volumes of “quality data” increases (companies as well as the NHS are increasingly “data-needy” from wanting to quickly develop working AI solutions), the risk of overlooking privacy concerns of the public in-

creases and can inflate a previous challenge mentioned earlier, which is building public confidence. A case in point is shown in an article in The Times at the end of 2019, where it was reported that Amazon was given free access to NHS data and that “The \$863bn company (Amazon) can access ‘all health-care information’ gathered by the NHS at the UK taxpayers’ expense, including ‘symptoms, causes and definitions’... Amazon can use the information to make, advertise and sell ‘new products, applications, cloud-based services and/or distributed software’ and can share it with third parties”. The Times also stated that “A commercial lawyer who analysed the contract said – ‘the most alarming thing is that Amazon isn’t paying anything for this and the data is very valuable’”. When the deal between the NHS and Amazon was first struck around July 2019, it raised many public concerns. Phil Booth, co-ordinator at medConfidential said: “personal and health data is heavily protected under GDPR in the UK, but Amazon Alexa doesn’t comply with the same laws”, with Big Brother Watch “labelling the partnership a ‘data protection disaster waiting to happen’”. This isn’t an isolated incident - as mentioned earlier, there are conclusions from the ICO that the “Google DeepMind NHS app test broke UK privacy law” and news stories showing NHS trusts giving Google the ability to process confidential patient data in combination with Google declining to release contract details. This lack of transparency will ultimately break public confidence and lead to a evaporation in trust between patients and the NHS, which would have several follow-on consequences, as well as killing the potential discussed up until this point.

Lastly, the issue of bias is introduced. As AI solutions are only as good as the data fed into them, we see countless examples of algorithms displaying existing inequalities and even exacerbating such

inequalities due to the data being used to train such AI solutions incorporating existing biases in the world today. One example of this having real life consequences is in America, where an algorithm used in many US hospitals was found “less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients with complex medical needs.” - introducing a two-tier system in healthcare and showing serious risk of perpetuating inequality to a level seen when racial segregation was prevalent in the United States. The data that is being used to train such algorithms must be closely examined and scrutinised before being used and even once a result is generated such results must be further dissected. The challenges around data privacy are discussed in Section 3: Conclusions: Policy Proposals.

Not only are individuals’ privacy at risk, but there is also a real risk of unmonitored AI solutions taking existing institutional racism within society and showing, perpetuating, and even expanding it. This means that safeguards must be put in place. Opponents to safeguards might call it “red tape”, “unnecessary bureaucracy” and say it slows down the implementation of AI solutions, but such measures being put in place will mean the NHS will make better use AI solutions quicker, as where public confidence supports it and individuals feel safe with its implementation the NHS will ultimately get much more support in the long-run.

We must keep in mind that it is the general public and NHS staff who will be working with AI solutions and who are the biggest stakeholders in all this. We need them to play a central role in the implementation of AI solutions in all areas of life, as well as health, if we are to fully realise the potential of AI within existing society. 🚩

CONCLUSIONS AND RECOMMENDATIONS

FOR AI IN HEALTHCARE

To bring everything discussed together, a strategy should be formulated to maximise potential. Going forward, to ensure that AI can be a success in healthcare, three T's must be satisfied, with trust being central to the strategy:

- **Transparency:** The way trust is built is through the public knowing exactly how their data is being processed, who is handling such data and what conclusions are being made from that data. They must also, in plain English, be informed on how their data is being used and in a language they understand, not told. We all have the experience of signing up a privacy agreement when accessing services, even though we don't know what we're agreeing to (simply to avoid reading the jargonised privacy policy). This is even more critical within healthcare, as the difference here is that we don't choose to access a service or not - we rely on health services, and patients have every right to know how their data is to be used. Through this transparency, other organisations and individuals such as publicly backed independent bodies, academics and elected officials can scrutinise how patient data is being used and ensure accountability is maintained and upheld.
- **Trust:** Being transparent, open and honest about how data is managed builds trust amongst healthcare staff and patients. From that, the public and healthcare professionals become more and more keen to see how AI can next improve healthcare and with more and more staff being more keen to make use of AI solutions in their practices.

- **Transition:** Once the public and healthcare professionals have overwhelming confidence in the use of AI in healthcare, we can move to how the healthcare system can be transformed - moving from reactive to preventive healthcare through use of data, where humans do what they do best and allowing AI to do what it does best. This is achieved through setting up NHS data to be 21st century ready (which will require world-class data management), making its data adaptable, high quality and fit for the future.

To execute such a strategy, several policy recommendations are presented below:

- A campaign or government programme aimed informing the public of their rights surrounding the handling their data as well as the relevant regulations
 - This is done in conjunction with robust regulation and governance to monitor and mitigate any inappropriate use of patient data
- The setting up of an independent body which inspects and stamp-marks companies and freelancers which meet the required standard in data handling and management, with the ability to rescind such accreditation
 - See Section 3: Conclusions: Policy Proposals for more details on a dedicated AI regulator and data standards
- A feasibility study should be enacted to look at legislation which requires data to be of a certain "quality" (including but not limited to no bias and exist-

ence of structural inequalities within society) before it can be used to train AI models which are to be used in the NHS


- Setting up of a patient working group as well as a staff working group on AI within the NHS: this would be comprised of a diverse number of patients and NHS staff and its goal would be to meet and discuss with NHS leaders their thoughts, ideas, concerns; and to represent the wide range of NHS patients and staff in how AI is implemented into the NHS, ensuring that the public and staff are brought along on the journey
- Ensuring recommendations made by AI solutions on healthcare is only provided to professionals once they have come to a conclusion about the appropriate course of healthcare for a patient themselves, to ensure AI solutions assist professionals rather than replacing them outright and to avoid anchoring bias
- The planning of NHS data to be streamlined and reorganised through world-class data management and restructuring such data in a way which makes it "21st century ready", allowing for better quality data to be used for when AI solutions are to be developed, either by the NHS and/or external companies, with the stipulation that the data is fully anonymised and handled with consent of the patients. This is achieved through setting up "processes and data workflows to aggregate, clean and convert these fragmented and isolated data records into a single high-quality, analysable data set"⁵³ such that we eliminate the fragmentation within NHS data, and

combine electronic health records all under one system to produce a high quality customer-centric dataset. This is to ensure sustainability, flexibility and “future-proofs” the data, such that there remains the ability to adapt to future changes in requirements and incoming breakthroughs

- Ability to incorporate data from other sources about a patient (with the patient’s permission), such as data from wearable technology to smart home devices, in combination with the data about a patient that is available in the NHS. Patients are then empowered to decide what data can be used to contribute to the outcome of the verdict of an AI model which makes up their healthcare (so long as a human professional remains in control and has the ability to overrule the verdict of an AI model). We must also keep in mind confirmation bias and the risk of “cherry picking” by patients, where the patient may choose data to obtain a certain outcome or to present an image of their health which isn’t real. Again, that is why a human professional must remain central to the providing of healthcare to a patient, as the human skills of a professional is what makes healthcare so effective. An example

where this can be displayed is where a patient is suffering from an addiction and decides to provide data which they may deem favourable to generate a favourable verdict from an AI model, which is not in the patient’s interest. This could end up with the patient missing out on life-saving treatment

- A push towards using open-source solutions or using existing open source frameworks in development
 - Improves the transparency of AI solutions that are implemented into the NHS
 - It would allow anyone to scrutinise how algorithms are coming to decision they are coming to, including academics, scrutiny panels and the public
 - Improves the security of AI solutions that are implemented into the NHS
 - Enables healthcare professionals to participate in the broader debate around AI, resulting in their understanding of such AI technology increasing. This leads to improved healthcare through allowing professionals to know how such AI technologies have come to a specific decision

- Requiring clinical trials on AI technology that are being planned to be used in healthcare to be made mandatory. This is to ensure their effectiveness in practice (currently, clinical trials aren’t required for AI use cases by NHS trusts)
- The setting up of a “data donor” scheme, such that individuals, like an organ donor scheme, can donate their data for research purposes and to develop further AI solutions. This should be started when the public is at a point where they feel comfortable in contributing to the implementation of AI within healthcare
 - This would give strict access for academics, companies, and freelancers (which are stamp-marked, as explained above) such that they can use the data to develop AI solutions
 - There is an ongoing discussion on whether citizens should be forced to give up their data if the benefits to the collective are great enough. Mandatory data collection would ensure that voluntary data is not skewed towards a certain demographic, which seems likely. This is discussed in the Data section of Section 3: Conclusions: Policy Proposals. 

References

1. Densen, Transactions of the American Clinical and Climatological Association, 48.
2. IBM Research, quoted in IBM Healthcare and Life Sciences, The future of health is cognitive, 3.
3. New Statesman, Matt Hancock unveils £250m fund for NHS AI.
4. Pulse, Health secretary launches £140m competition for NHS artificial intelligence projects.
5. Patel et al., Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods, quoted in ScienceDaily, Artificial intelligence expedites breast cancer risk prediction.
6. Incisive Health, Saving lives, averting costs, 7.
7. Office of National Statistics, "Avoidable mortality in the UK 2016", quoted in Macmillan Cancer Support, Statistics fact sheet, 9.
8. World Health Organization, Cancer, Key statistics.
9. World Health Organization, Cancer, Cancer Prevention.
10. Cancer Research UK, Why is early diagnosis important?.
11. BBC News, 'Tens of thousands' of heart procedures delayed by pandemic.
12. Cancer Research UK, Over 2 million people in backlog for cancer care.
13. BBC News, Coronavirus: NHS waiting list 'could hit 10 million this year'
14. McKinney et al., International evaluation of an AI system for breast cancer screening, 89.
15. The Royal College of Radiologists, The NHS does not have enough radiologists to keep patients safe.
16. FitBit, Privacy Policy.
17. Stat, IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show.
18. Stone et al., Artificial Intelligence and Life in 2030. Stanford, CA: Stanford University, September 2016.
19. National Health Service, Acute kidney injury.
20. Kidney Care UK, Facts and Stats.
21. National Confidential Enquiry into Patient Outcome and Death, Adding Insult to Injury, 40.
22. Tomašev et al., A clinically applicable approach to continuous prediction of future acute kidney injury, 116
23. Tomašev et al., 116
24. Wired, DeepMind's new AI predicts kidney injury two days before it happens.
25. Tomašev et al., 118, fig. 3.
26. Wired, DeepMind's new AI predicts kidney injury two days before it happens.
27. Digital Health, DeepMind's Streams app saves £2,000 per patient, peer review finds.
28. Telegraph, Google's DeepMind can now predict deadly kidney injury two days before it happens.
29. New Scientist, It's too soon to tell if DeepMind's medical AI will save any lives.
30. Science Media Centre, Expert reaction to study on a deep learning approach to predicting acute kidney injury.
31. Science Media Centre, Expert reaction to study on a deep learning approach to predicting acute kidney injury.
32. LeadingAge, Cerner, Keystone Technologies, Stanley Healthcare, and MatrixCare Make Summer News.
33. Food and Drug Administration, FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures.
34. The Hill, FDA chief moves to promote artificial intelligence in health care.
35. Davies (the Comptroller and Auditor General), "NHS financial management and sustainability".
36. Institute for Fiscal Studies, Securing the future: funding health and social care to the 2030s, quoted in The Kings Fund, The NHS budget and how it has changed.
37. Birkhäuser et al., Trust in the health care professional and health outcome: A meta-analysis.
38. YouGov, NHS Fieldwork Dates: 11th - 12th March 2019, quoted in Sensyne Health, Survey shows strong support for analysis of anonymised NHS patient data for medical research.
39. YouGov, Sample size: 1,027 healthcare professionals in the UK Fieldwork: 23 October - 31 October 2019, quoted in Sensyne Health, NHS staff fears over multinational 'big tech' firms analysing patient data.
40. Invoca, New Invoca Research Conducted by The Harris Poll, quoted in HIT Consultant Media, Only 20% of Consumers Would Trust AI-Generated Advice for Healthcare.
41. Information Commissioner's Office, Royal Free - Google DeepMind trial failed to comply with data protection law.
42. Politics Home, The public need confidence that the Contact Tracing App will protect their privacy.
43. Royal College of General Practitioners, Apps and algorithms may 'support but will never replace' GPs.
44. Barton-Laws et al., Factors associated with patient recall of key information in ambulatory specialty care visits: Results of an innovative methodology, 7.
45. Ernst & Young, Realising the value of health care data: a framework for the future, 1
46. Ernst & Young, 20
47. Ernst & Young, 12, fig. 4.
48. The Times, Amazon ready to cash in on free access to NHS data.
49. Digital Health, Amazon Alexa partnership puts patient data 'at risk'.
50. BBC News, Google DeepMind NHS app test broke UK privacy law.
51. i News, NHS trusts give Google green light to process confidential patient data.
52. Nature, Millions of black people affected by racial bias in health-care algorithms.
53. Ernst & Young, 11

Bibliography

Barton Laws, Martin, Yoojin Lee, Tatiana Taubin, William H. Rogers, and Ira B. Wilson. "Factors associated with patient recall of key information in ambulatory specialty care visits: Results of an innovative methodology". PLOS ONE 13(2): e0191940. (2018) doi: <https://doi.org/10.1371/journal.pone.0191940>

BBC News. "Google DeepMind NHS app test broke UK privacy law". July 3, 2017. <https://www.bbc.co.uk/news/technology-40483202>

Birkhäuser, Johanna, Jens Gaab, Joe Kossowsky, Sebastian Hasler, Peter Krummenacher, Christoph Werner, and Heike Gerger. "Trust in the health care professional and health outcome: A meta-analysis." PLoS ONE 12, no. 2 (2017): e0170988. Gale OneFile: Health and Medicine.

Brian Ferguson, "Investing in prevention: the need to make the case now", Public Health Matters, February 22, 2016. <https://publichealthmatters.blog.gov.uk/2016/02/22/investing-in-prevention-the-need-to-make-the-case-now/>

Cancer Research UK. "Over 2 million people in backlog for cancer care". June, 1, 2020. <https://www.cancerresearchuk.org/about-us/cancer-news/press-release/2020-06-01-over-2-million-people-in-backlog-for-cancer-care>

Cancer Research UK. "Why is early diagnosis important?". June, 26, 2018. <https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important>

Digital Health. "Amazon Alexa partnership puts patient data 'at risk'". July 10, 2019. <https://www.digitalhealth.net/2019/07/amazon-alexa-nhs-patient-data-safety/>

Digital Health. "DeepMind's Streams app saves £2,000 per patient, peer review finds". July, 31, 2019. <https://www.digitalhealth.net/2019/07/deepminds-streams-saves-2000-peer-review/>

Densen, Peter. Challenges and opportunities facing medical education. Transactions of the American Clinical and Climatological Association 122, (2011): 48–58.

Ernst & Young. "Realising the value of health care data: a framework for the future". 2019.

FitBit. "FitBit Privacy Policy". Effective: December 18, 2019. <https://www.fitbit.com/us/legal/privacy-policy>

Food and Drug Administration, 2018, "FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures" <https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures>

Gareth Davies (the Comptroller and Auditor General), "NHS financial management and sustainability". National Audit Office, 2020: 12 <https://www.nao.org.uk/report/nhs-financial-management-and-sustainability/>

HIT Consultant Media. "Only 20% of Consumers Would Trust AI-Generated Advice for Healthcare". July, 1, 2019. <https://hitconsultant.net/2019/07/01/survey-only-20-of-consumers-would-trust-ai-generated-advice-for-healthcare/#.XvCw3mhKhPa>

i New. "NHS trusts give Google green light to process confidential patient data". September 18, 2019. <https://inews.co.uk/news/nhs-patient-data-google-privacy-634080>

IBM Research, The future of health is cognitive. United States of America: IBM, 2016.

Incisive Health. Saving lives, averting costs An analysis of the financial implications of achieving earlier

diagnosis of colorectal, lung and ovarian cancer. Cancer Research UK, 7.

Information Commissioner's Office, "Royal Free - Google

DeepMind trial failed to comply with data protection law". July 3, 2017. <https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/>

ITV. "Warnings some local authorities could go bust the longer coronavirus crisis continues". April 2, 2020. <https://www.itv.com/news/calendar/2020-04-02/warnings-some-local-authorities-could-go-bust-the-longer-the-coronavirus-crisis-continues/>

Kidney Care UK. "Facts and Stats". <https://www.kidney-careuk.org/news-and-campaigns/facts-and-stats/>

LeadingAge. "Cerner, Keystone Technologies, Stanley Healthcare, and MatrixCare Make Summer News". August 18, 2015. <https://www.leadingage.org/cerner-keystone-technologies-stanley-healthcare-and-matrix-care-make-summer-news>

Macmillan Cancer Support. "Statistics fact sheet". February 2019. https://www.macmillan.org.uk/_images/cancer-statistics-factsheet_tcm9-260514.pdf

McKinney, S.M., Sieniek, M., Godbole, V. et al. "International evaluation of an AI system for breast cancer screening". Nature 577, (2020): 89–94. doi: <https://doi.org/10.1038/s41586-019-1799-6>

National Confidential Enquiry into Patient Outcome and Death (NCEPOD). "Adding Insult to Injury: A review of the care of patients who died in hospital with a primary diagnosis of acute kidney injury (acute renal failure)". London: NCEPOD, 2009. https://www.ncepod.org.uk/2009report/Downloads/AKI_report.pdf

National Health Service. "Acute kidney injury". February, 25, 2019. <https://www.nhs.uk/conditions/acute-kidney-injury/>

Nature. "Millions of black people affected by racial bias in health-care algorithms". October 24, 2019. <https://>



www.nature.com/articles/d41586-019-03228-6

New Scientist. "It's too soon to tell if DeepMind's medical AI will save any lives". July 31, 2019. <https://www.newscientist.com/article/2212100-its-too-soon-to-tell-if-deepminds-medical-ai-will-save-any-lives/#ix-zz6Q1c0SB9a>

New Statesman. Matt Hancock unveils £250m fund for NHS AI. August 8th 2019. <https://tech.newstatesman.com/policy/matt-hancock-nhs-ai>

Patel, Tejal, Mamta Puppala, Richard Ogunti, Joe Ensor, Tiancheng He, Jitesh Shewale, Donna P Ankerst, Virginia Kakiamani, Angel Rodriguez, Stephen Wong and Jenny Chang. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. *Cancer* vol. 123,1 (2017): 114-121.

Politics Home. "The public need confidence that the Contact Tracing App will protect their privacy". 19 May 2020. <https://www.politicshome.com/thehouse/article/the-public-need-confidence-that-the-contact-tracing-app-will-protect-their-privac>

Pulse. Health secretary launches £140m competition for NHS artificial intelligence projects. January 29th 2020. <http://www.pulsetoday.co.uk/news/health-secretary-launches-140m-competition-for-nhs-artificial-intelligence-projects/20040049.article>

Royal College of General Practitioners. "Apps and algorithms may 'support but will never replace' GPs, says RCGP". June 27, 2018. <https://www.rcgp.org.uk/about-us/news/2018/june/apps-and-algorithms-may-support-but-will-never-replace-gps-says-rcgp.aspx>

ScienceDaily. "Artificial intelligence expedites breast cancer risk prediction". August, 29, 2016. www.sciencedaily.com/releases/2016/08/160829122106.htm

Science Media Centre. "Expert reaction to study on a deep learning approach to predicting acute kidney injury". July, 31, 2019. <https://www.sciencemediacentre.org/expert-reaction-to-study-on-a-deep-learning-approach-to-predicting-acute-kidney-injury/>

Sensyne Health. "NHS staff fears over multinational 'big tech' firms analysing patient data" November 23, 2019 <https://www.sensynehealth.com/news/nhs-staff-fears-over-multinational-big-tech-firms-analysing-patient-data>

Sensyne Health. "Survey shows strong support for analysis of anonymised NHS patient data for medical research". June, 8, 2019 <https://www.sensynehealth.com/newsroom/yougov-uk-public-support-analysis-anonymised-nhs-patient-data>

Stat. "IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show". July 25, 2018. <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/>

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*. Stanford, CA: Stanford University, September 2016. <http://ai100.stanford.edu/2016-report>

Telegraph. "Google's DeepMind can now predict deadly kidney injury two days before it happens". July, 31, 2019. <https://www.telegraph.co.uk/science/2019/07/31/googles-deepmind-can-now-predict-deadly-kidney-injury-two-days/>

[jury-two-days/](#)

The Hill. "FDA chief moves to promote artificial intelligence in health care". April 26, 2018. <https://thehill.com/policy/healthcare/385020-fda-chief-moves-to-promote-artificial-intelligence-in-health-care>

The Kings Fund. "The NHS budget and how it has changed". March, 13, 2020. <https://www.kingsfund.org.uk/projects/nhs-in-a-nutshell/nhs-budget>

The Royal College of Radiologists. "The NHS does not have enough radiologists to keep patients safe, say three-in-four hospital imaging bosses". April, 4, 2019. <https://www.rcr.ac.uk/posts/nhs-does-not-have-enough-radiologists-keep-patients-safe-say-three-four-hospital-imaging>

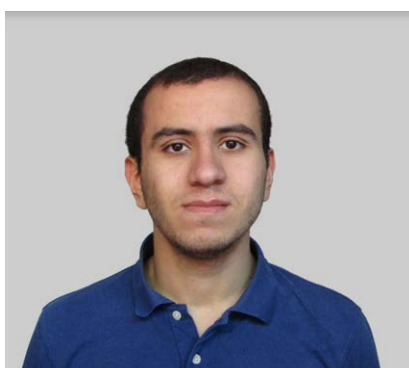
The Times. "Amazon ready to cash in on free access to NHS data". December 8, 2019. <https://www.thetimes.co.uk/article/amazon-ready-to-cash-in-on-free-access-to-nhs-data-bbzp52n5m>

Tomašev, N., Glorot, X., Rae, J.W. et al. "A clinically applicable approach to continuous prediction of future acute kidney injury". *Nature* 572, (2019): 116–119. doi: <https://doi.org/10.1038/s41586-019-1390-1>

Wired. "DeepMind's new AI predicts kidney injury two days before it happens". July, 31, 2019. <https://www.wired.co.uk/article/deepmind-streams-ai-algorithm-kidney-injury>

World Health Organization. "Cancer, Cancer Prevention". <https://www.who.int/cancer/prevention/en/>

World Health Organization. "Cancer, Key statistics". <https://www.who.int/cancer/resources/keyfacts/en/>



Mohamed Hameda is an undergraduate student at the University of St Andrews reading Computer Science and Mathematics with a focus on Statistics. During the pandemic, he has been contributing to Scientists for Labour and its COVID-19 briefings and reports, bringing the latest research on COVID-19 to Labour Party politicians and advisers. Mohamed is also a member of the Co-op National Members Council, a charity trustee and has previously worked at the Cabinet Office.

AUTOMATING DEFENCE: INNOVATION AND THE INCREASING ROLE OF ARTIFICIAL INTELLIGENCE IN WARFARE AND NATIONAL SECURITY

By Luke Richards

“Whoever leads in AI will rule the world.” – Vladimir Putin

“By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it.” – Eliezer Yudkowsky

The wars of the 21st century are going to become increasingly nasty, brutish and short. Their ferocity and intensity are developing in tandem with new technologies and methods of warfare. Meanwhile, the development of AI technologies opens up a possible future of a world radically transformed, and it is no less so within the realm of warfare and national security.

The above oft-quoted statement by the Russian President Vladimir Putin only tells part of the story about the role of AI in the future, and those who look to wield its power. Foremost, AI is only as powerful as the computer systems it runs on, the information it is fed, and the complex international innovation systems that make its various parts. This detailed web of socio-technical systems that underpin the emergence and innovation of new and existing technologies are increasingly found in national security debates. It is a situation that has seen an increasing politicisation as grey zone activities, in which battles are often fought at

the sub-threshold of war, become the norm.

Over 200 years have elapsed since the military theorist Carl von Clausewitz defined war as “the continuation of politics by other means”. Much has changed since, and drastic transformation is on the horizon as mechanisation becomes intertwined with machine intelligence. Yet, technology too is deeply political - and its development is not linear. That is, we cannot expect technology, such as AI, to develop unless we make active choices as societies towards doing so, and our actions determine the outcome. The technological tools and methods of war are then themselves a continuation of politics by other means.

For some, defence spending can be considered a public good, but for others, it can also contribute to global insecurity. Technologies are said to have a dual-use when they have both civilian and military applications. While the majority of AI innovation is currently found in the private sector for civilian purposes some of these developments are

likely to have a dual-use. In addition to this, it looks as if AI will become a general-purpose technology and be diffused widely across society within a broad spectrum of applications. Given this, it can be difficult to prise apart the civilian and military aspects of AI development, especially when considering second order impacts and innovations. Something made more difficult since innovation does not take place in a vacuum, it does so alongside and within other states, and their competing notions of security and insecurity.

Due to the intermingling and blurring of the civil-military dichotomy, it will be useful to briefly outline the relationship of innovation with national security before coming addressing AI technologies and applications more directly. The themes within this chapter outline how the state can be proactive, adaptive and even entrepreneurial in developing AI and other emerging technologies, and how it won't be able to just rely on the market in the future.



AI, innovation, and national security

Historically the defence sector has played a nurturing role in the development of numerous technologies. The internet being one example, its historical lineage is rooted in the creation of the military network ARPANET, which inspired technologies that now underpin the net as we know it. Gov-

ernment spending on defence is one of the largest contributors to research and development (R&D) globally. Comparatively, the UK has sold off active parts of the state under the mantra of ‘the private sector knows better’ since the 1980s. Yet, the defence sector, and all its industrial might, has often

been bolstered by state support. Advances in AI, however, have often come from commercial sector innovation, and defence has been left trying to catch up. This situation has forced the defence sector to lean more on commercial enterprises for access to technologies more so than it may have done in

the past.

National security concerns may drive Britain, and other states, to change the way they currently approach technology and trade. Recently, the increasing politicisation of technological development has made its way into public life through increasingly bitter spats between the US and China, often referred to as the tech war. Legitimate security concerns have become mired in other geopolitical and economic ambitions and agendas. Debates around what technologies and where they come from and who they are sold to have historically stayed far from public discourse. However, as new technologies allow our everyday lives to become weaponised, new approaches towards innovation, governance and regulation will have to be taken.

For years the UK has failed to invest in R&D, and has shied away from developing a comprehensive industrial strategy. Some have called for the creation of a British version of the United States' Defence Advanced Research Project

AI in defence

The use of AI within defence is both a pressing issue and one that is subject to much hype. As with most facets of AI, there is an array of potential applications to its use and both incremental and radical changes across this broad spectrum of applications is possible.

AI technologies are likely to be deployed across a range of cyber electromagnetic activities, acting across both the physical and cognitive domains. It could also be deployed in a more direct role as a force enabler, through the use of increasingly automated weapons systems, for example. The time horizon for the development of artificial general intelligence is unknown or, indeed, not possible to predict. Instead, human-machine

Agency to drive new innovation and fund seemingly wacky ideas. If implemented correctly, it could go some way to remedy previous policy deficiencies, although a plethora of challenges will remain. For now, the MoD is attempting to drive innovation around AI technologies through the private sector and academia. Projects such as the Defence and Security Accelerator and the National Cyber Security Centre's Cyber Accelerator are examples of the state attempting to ferment innovation within the private industry. Schemes such as this can create both the skills the country needs within its workforce and an industrial base it can draw from for the sake of its security.

Britain will not be able to develop entirely new industries in every area. In some areas, it will have to rely on the industrial skills of those it trusts for specific products. Conversely, it may also have to think more about the products British companies make, and the markets they sell to. Such changes will challenge prevalent notions of free trade.

teaming using narrow AI systems complementing human intelligence is much more likely.

With a rise in automated and increasingly digital systems will also come a compression in time, as the tempo and speed in which war takes place increases. The full-spectrum digital battlefield will produce volumes of data beyond the scope of human cognition. So, AI will be needed just to keep up with what is taking place. Changes brought about by AI-powered systems within the realm of defence could destabilise existing norms of warfare and create new problems within existing security issues, such as nuclear.

The issue that perhaps gets the most attention is the development of lethal autonomous weapons

Defence innovation has shown how an active state can create an environment for clusters of cutting edge industries to grow. There is little reason why the innovation measures being fostered for defence could not be applied elsewhere - for say, helping solve societal grand-challenges, such as global warming. A co-production of tech development between the state and industry in other areas would also allow the state to build relevant technical skills for civic purposes and institutional memory to draw from to better regulate.

This section has outlined how the role of innovation, technologies such as AI, and national security concerns are likely to become more prominent and the politics around it fraught. Yet such technology is being developed, and in some areas Britain is leading, but it comes with risks and opportunities. The rest of this chapter will explore the use of AI in defence.



systems (LAWS). There is an array of ethical and legal issues that have arisen through debates on whether autonomous machines that can kill should even be developed. There is also some debate over what a lethal autonomous weapons system, as opposed to an automated weapons system, is. The MoD makes a clear distinction between the two:

Automated: "Response to inputs from one or more sensors, is programmed to logically follow a predefined set of rules in order to provide an outcome. Knowing the set of rules under which it is operating means that its output is predictable."

Autonomous: "is capable of understanding higher-level intent and direction. From this understanding

and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.”

Britain is active in multilateral fora and global dialogue discussing the issue of LAWS. The main focus of the debate so far has been the Convention on Certain Conventional Weapons (CCW).


In 2016 the UK government stated that “the UK believes that LAWS do not, and may never, exist. Furthermore, we have no intention of ever developing systems that could op-

erate without any human control.” At present, Britain is one of only 12 countries to come out against a preemptive ban on such systems.

Human control has become a central feature to debates at the international level. It places human perception and judgment on whether to use lethal force in a specific instance at its centre.

The UK’s stance towards LAWS is centred on allowing for its development, but with a focus on humans ultimately being responsible for machines and their actions. In such a scenario, weapons could be powered by varying degrees of autonomy depending on the context they’re used. Many weapons systems already have a degree of autonomy; a missile once fired will navigate towards its target, for example. There is a spectrum of

autonomy that can be achieved within weapons systems. It encompasses the capacity for machines to assess their surroundings and what to strike.

Discussing the use of AI in defence needs to be broader than LAWS as the technology advances and it finds purpose in a plethora of other ways. The first waves of automation within the defence sector are already happening, and relatively inane compared to the fear aroused over LAWS. AI is currently applied within areas such as logistics, geospatial analysis, intelligence and cybersecurity to aid humans. AI and automation are becoming more profound and applied without any controversy; warfare and defence is already becoming automated - the consequences of which it is too early to know. 

AI, politics and technological futures

The politics of technology is often never at the forefront of either its development or use. This has shifted recently as concerns around the potential security issues around technology have become more prominent in the civilian sphere as states in the West have begun to question China’s international ambitions. The recent Huawei debate in Britain has displayed if anything that Britain simply doesn’t have the manufacturing capability or the knowhow on a national scale to produce certain technologies domestically and is reliant on others.

It is often tricky to forecast technological change, and perhaps even more so alongside its relation to national security. When it comes to, say, 5G, Britain is entirely reliant on a small number of equipment manufacturers. Issues such as this also fall over into the arena of AI. For example, countries are dependent on a handful of states that can produce the most advanced computer chips. Those that can are further dependent on states that manufac-

ture lithography equipment within their plants. Britain cannot create, and it is unlikely to develop the ability to manufacture such items. For all the country’s political ambitions to enhance aspects of what it perceives as its sovereignty, it will remain technologically dependent on others. Global politics and technological futures will have to be addressed going forward at fundamental policy and regulatory levels. Especially as AI development within the UK will be subject to international pressures and unforeseen events.

The European Union (EU) is attempting to develop parts of its tech sector to achieve technological sovereignty. It is an attempt to make it neither over-reliant on the US or China for the supply of specific technologies. Britain may be a world leader in some areas of AI development. Still, it is entirely dependent on others to build the hardware systems the software runs on. Compared to the EU, Britain has neither the economic or normative power to shift the global

development of technologies in its favour.

When Britain leaves the EU and takes up its mantle as a ‘global free-trading nation’ it may have to make increasingly stark choices around trade and national security. It is likely to find itself stuck between the US and China over technologies and may be forced to take one side or another. Given this, technological insecurity could become an increasing worry for Britain’s national security.

The developing civil-military nexus and widening security paradigms of AI, and many other technologies, will raise issues for all subsequent British governments. Security concerns will need a more proactive and reactive state, one that will have to work closely with the private sector. However, Britain’s ability to govern and direct the course of the development of AI and other emerging technologies in general could find itself diminished somewhat as time goes on. That is not to say it should not try and align itself with a version of the future that

is both ethical and democratic. It should attempt to create this future through partnerships with its allies and the wider international community and push to develop AI for good. 🇬🇧

NATIONAL DEFENCE DIGITALISATION: CYBER SECURITY AND ARTIFICIAL INTELLIGENCE

Cyber is one area where the UK punches above its weight globally. Britain has a robust cybersecurity sector and has implemented a whole of society approach to dealing with cyber risks.

In addition to this, it has a well-connected cyber-ecosystem and world-leading expertise within the government in organisations such as Government Communications Headquarters (GCHQ) and its civilian oriented arm the National Cyber Security Centre (NCSC). The country's cyber mass is also bolstered by specialists within Defence Intelligence and various military units. As with AI, cyber capabilities, skills and innovations are concentrated within the private sector. This situation has seen the defence sector encourage innovation for national security. Going forward, the development of AI will bring challenges and benefits to cybersecurity.

The perils and pitfalls of cyber and AI are entangled with the broader digitalisation of society. Cyber technologies have allowed for an increasingly interconnected world with seemingly amorphous boundaries in which a multitude of state and non-state actors operate. Advances in AI have been driven by the reams of data that digital technologies can increasingly capture; especially as variants of cyberspace have become ever more prevalent in everyday life. The ubiquity of digital technology Cyber is one area where the UK punches above its weight globally. Britain has a robust cybersecurity sector and has implemented a whole of society approach to dealing with cyber risks.

In addition to this, it has a well-connected cyber-ecosystem and world-leading expertise within the government in organisations such

as Government Communications Headquarters (GCHQ) and its civilian oriented arm the National Cyber Security Centre (NCSC). The country's cyber mass is also bolstered by specialists within Defence Intelligence and various military units. As with AI, cyber capabilities, skills and innovations are concentrated within the private sector. This situation has seen the defence sector encourage innovation for national security. Going forward, the development of AI will bring challenges and benefits to cybersecurity.

The perils and pitfalls of cyber and AI are entangled with the broader digitalisation of society. Cyber technologies have allowed for an increasingly interconnected world with seemingly amorphous boundaries in which a multitude of state and non-state actors operate. Advances in AI have been driven by the reams of data that digital technologies can increasingly capture; especially as variants of cyberspace have become ever more prevalent in everyday life. The ubiquity of digital technology and its ability to transgress national borders has created unique and unforeseen vectors for attack and disruption.

Ongoing efforts have been made to secure the digital realm. The role of the state has been marked by the degree of fluidity it has needed to adapt to cyber challenges. For example, the UK's first National Cybersecurity Strategy was released in 2011 and looked to the market to drive changes. However, the latest (2018) notes that "this approach has not achieved the scale and pace of change required to stay ahead of the fast-moving threat." The UK government has had to intervene and actively invest and support market forces to build its

cyber competencies.

What has developed from the government's cyber policies looks like what the economist Mariana Mazzucato might call the entrepreneurial state. Not only has the government intervened across society to implement its strategy, but it has also actively pursued innovation-led growth using both state and market forces. In doing so, it has encouraged business start-ups around clusters of cyber businesses and skills within the UK to create a level of cyber self-sufficiency. In some areas of national security the state is partly reliant on the private sector. The UK military's Joint Cyber Reserve Force is one example of one way in which the military has reached out to those in the private sector with the skills it needs. So the UK's cyber industrial base, supported in part by the state, gives it a broader well of expertise to draw from.

Cyber technologies are a conduit through which AI can operate, and the two areas are closely linked, as shown by the examples below. 🇬🇧

AI and networked warfare

Networked computer systems are vulnerable to exploitation, and AI systems are being developed for offensive and defensive purposes.

There are many ways in which AI could be used for offensive cyber such as finding vulnerabilities in computer systems and propagating malicious code. On the defensive side, machine learning can help in sifting through data to detect abnormalities and potential cyber-attacks before they can cause damage. There is some debate around what 'cyberwar' is and if it is even possible. The advent of AI cyber technologies and increased automation could make it easier for states and non-state actors to develop weapons and target systems. However, computer systems are just a large part of a socio-technical cyber system that could be exploited or manipulated.

Data is becoming increasingly crucial to the way the military operates. Major General Tom Copinger-Symes CBE, Director of Military Digitisation, UK Strategic Command, notes that a blizzard of data is hitting our leaders and that we aren't making sufficient sense of the data and are at risk of drowning in the data and being asphyxiated by lack of understanding.

The military is increasingly reliant on both the production and transmission of information for data-centric informatized warfare: "Information is no longer just an enabler, it is a fully-fledged national lever of power, a critical enabler to understanding, decision-making and tempo, and a 'weapon' to be used from strategic to tactical levels for advantage." Big data is an enabling technology for AI, but conversely, AI is also going to be needed to sift through the increasing volume of

data future battlefields are going to generate and the potential strategic advantages found within it.

As more of the battlefield is digitalised, it also opens up new ways through which can be fought. A pertinent example of this would be hybrid warfare'. This method of warfare uses "the synchronised use of multiple instruments of power tailored to specific vulnerabilities across the full spectrum of societal functions to achieve synergistic effects." The UK, too, is looking to exploit data for information advantage and cognitive effects for its advantage. This method of warfare blurs distinctions between civilian and military elements. AI could enable the development of tools that could collect surveillance and information on large groups of civilians and military personnel. 🚩

Automation propaganda (see also: *Section 1: Communications*)

One of the ways AI could be utilised for hybrid warfare is the propagation of so-called fake news and the use of deep fakes.

A synthetic media created by AI tools - to spread propaganda or sow social dissent; a theme picked up by Tom Ascott in the following chapter. Digital technologies and products such as social media allow for both the dissemination of false narratives. It can also aid the collection of personal information that could be used to personalise propaganda and create alternative realities. The promulgation of synthetic media across cyber domains such as social media and messaging apps has the potential to cause social disruption and violent social unrest. It could rise to become a threat to democracy.

One of the highest-profile cases of automated propaganda is perhaps the Cambridge Analytica (CA) scandal. It involved the misuse of

data taken from Facebook to target and influence voters in various elections. The company built an AI-based system that automatically tested variations of an ad before deciding which one to place and combined it with personal data to target and sway groups of voters effectively. The increasing amount of data about individuals openly collected and available to various actors combined with new tools and techniques of persuasion, such as deepfakes, open up an avenue for our actions and data to be weaponised against us. The regulation of data and preventing data misuse could find itself crossing both civil and military boundaries.

The private sector is starting to respond to platforms being used for such purposes. For example, Facebook has recently created a machine learning model it calls Deep Entity Classification (DEC) to detect fake accounts by analysing the way in which they interact and whether

they use AI-generated profile images. The scale and scope of the tech giants and regulating them could become tricky from the perspective of international security. There is a tradeoff between over-regulating and hampering innovation if there are no clear international guidelines for doing. States could gain both a competitive and strategic advantage through not breaking up or regulating large firms. There are also tradeoffs between security and liberty when it comes to governing the use of cyber and data that won't be the same across states either. The EU is attempting to set normative standards in such areas and leading on user privacy through regulations such as GDPR. Britain's future regulatory regime will have to consider these dynamics.

Increased levels of competition at the sub-threshold of war and within the public sphere shows how disconnected this digital public space

can be from territorial boundaries. The cyber domain allows for entire realities to be created and false narratives normalised in the interests of one state over another. Yet the networks themselves are entirely physical. Looking ahead,

Britain should strive for a truly international approach to regulate AI, data, and related cyber technologies, especially as this is now a world in which AI can operate and act at scales and speed far beyond human comprehension. There is a

need to cast aside certain staid notions of sovereignty and give way to the realities of a deeply interconnected world to be truly effective. 🇷🇺



Luke works at the intersection of science and technology policy and international relations. He has worked on diverse issues such as AI policy and regulation, responsible innovation, and assessing the cyber capabilities of states. More recently, he has focussed on the relationship between state and private power in the development, diffusion and use of emerging technologies alongside ways to democratise science and technology to co-produce a future that works for everyone.

AI AND DISINFORMATION, AN ALGORITHMIC ASSAULT ON DEMOCRACY

By Tom Ascott

Disinformation is already altering our political landscape

Disinformation has already helped to shape more of our significant political choices than one would like to admit, and the consequences of such a sharp rise in information warfare campaigns are only starting to be fully understood.

Disinformation has played a role in not only recent UK elections, but the advancement of AI-driven disinformation technologies have worrying consequences for developing countries. Recently, an AI-created fake video sparked a failed military coup in Gabon. This technology has the extremely worrying potential to change the political landscape anywhere in the world.

This paper follows the EU Commission's High Level Expert Group on Fake News and Online Disinformation definition of disinformation as 'false, inaccurate, or misleading

information designed, presented and promoted to intentionally cause public harm or for profit'.¹

Disinformation, then, is not news stories that are found to be distasteful or disagreeable. Nor is it information that is peddled by self-styled experts or theorists – information that, while untrue, they believe to be true, which is misinformation.

In this context, information warfare is a way to disrupt an adversary from being able to collect, process, and disseminate information. Disinformation campaigns are used in information warfare to meddle with what people think, and to manipulate their opinions.

The longer that disinformation can persist, the more of a problem it creates. It only works because it exploits a simple but core democratic notion; that what one reads

online can be trusted to be true. This concept is what has allowed the Wiki foundation to flourish.

Fundamentally, disinformation corrupts the well of human knowledge – and it is not slowing down. Disinformation campaigns can currently prove beneficial for social media platforms, as they benefit from unclear and lax rules that allow for opaque political advertising. Such platforms build up detailed user profiles using thousands of data points. Even if these users remain anonymous to advertisers, their data is so comprehensive that adverts, or disinformation, can be tailored so specifically as to be incredibly convincing. Fears that advertising marketing and social media data are enough to unmask users are also deeply concerning.


Disinformation predates AI

This is the information warfare arena into which disinformation is currently being deployed, in order to move the pendulum on key strategic decisions through the manipulation of a trusting online culture and either using paid or organic groups.

The disinformation itself is made manually and is posted online by trolls. In his 2016 report, Special Counsel Robert S. Mueller III, defined a troll as a user that will 'post inflammatory or otherwise disrupt

ive content on social media or other websites'². For the most part, disinformation is still produced in 'troll farms' – offices where real people clock in every day, sit down at a computer and write disinformation online as part of a coordinated campaign³. These 'trolls' are paid and treat this like a normal job. Any website that has the ability for users to submit content or comments is a site where they can spread disinformation.

There are two inefficiencies in the

way that disinformation is currently being produced: quality and speed. In 2016, researchers from ZeroFOX⁴ tested SNAP_R (Social Network Automated Phishing with Reconnaissance) and found that it was six times faster at finding and engaging targets on Twitter, and five times more effective at converting them to click on malicious links, as compared to a human counterpart. 

AI can supercharge disinformation

AI can make great leaps in eradicating these inefficiencies, thus becoming a faster and more efficient tool for spreading disinformation.

The AI-powered algorithms used for targeting are already sophisticated. The online tools available to advertisers to find users and market to them are the same as those used in information warfare. In this instance, however, the bots themselves do not have to be particularly smart. A bot, or a software-controlled account, range from the simplest of designs to more sophisticated ones.

There is a low marginal cost to

having more bots on a network. At present, bots spreading disinformation do not have to be sophisticated because the disinformation itself isn't. The future, however, may start to present more advanced forms of disinformation which are increasingly in tune with individual user data for content tailoring. Deepak Dutt, the CEO of mobile security company Zighra, opined that AI will be used to 'mine large amounts of public domain and social network data to extract personally identifiable information like date of birth, gender, location, telephone numbers, e-mail addresses, and so on'⁵. This information can then be analysed by AI tools to cre-

ate disinformation that is tailored to individuals.

Such an approach would be effective as the psychological impact of disinformation is bolstered through repetition. The more a fine-tuned statement is repeated, the more likely that social network users will believe that it is true. This is called the 'illusory truth effect'⁶. An effective way to spread disinformation – and to ensure users' repeated exposure to it – is by message boosting. Bots do not need to send new and unique pieces of disinformation; instead, they can simply retweet or share existing pieces of disinformation that fit the same narrative. 🇷🇺

Deepfakes and disinformation

In a simplistic way, disinformation campaigns can be split into two basic projects; the generation of content which is intended to manipulate an audience, and an ability to distribute that content.

AI will greatly improve the ability for disinformation campaigns to distribute content, but it also provides for a greater ability to create content for those campaigns. The future of this content are deepfakes.

Deepfakes are videos that are made by AI and replace the face, and sometimes voice, of one person with another. They are highly realistic⁷, and can be easier to make than any previous form of video editing or manipulation.

Currently, the capacity of trolls who wage information warfare is limited by their ability to create visual propaganda. This includes shoddily captioned memes or photo-shopped images that are often of a very low quality. Memes can be thought of captioned images one might encounter on social media sites - for example 'Pepe the Frog', the green cartoon frog that is often

seen online.

If trolls are unable to create either of those then they will have to write text posts. Russian trolls can be betrayed by spelling, grammar or taxonomic errors⁸. However, deepfakes offer the potential for AI to generate the propaganda for them.

The spread of memes and meme culture gives an insight into how one might anticipate deepfakes to spread. Memes have come from online image boards such as 4chan, to the mainstream websites, such as Facebook, Instagram or Twitter. Presently, it is easy to find deepfakes on image boards that depict unethical content. 96% of deepfakes are of non-consensual pornography⁹. Unless social media sites intervene, then it is only a matter of time until more of this type of content will be easily found on mainstream sites, much in the same way memes from image boards have become a staple of social media sites.

Deepfakes will only get better over time, becoming more convincing to the human eye, using less footage to be made, and will be fast-

er and cheaper to produce. They are yet to reach their full potential, and there is plenty of private funding that is interested in advancing this technology. As they get easier to make their different applications will be better understood by each sector, and they have already made their way into politics.

In Gabon, a deepfake has already inspired a failed coup. President Ali Bongo left Gabon after suffering from a stroke. Months later, and after rumours of his death had started to circulate, the Vice President announced that President Bongo had suffered a stroke. An alleged deepfake video was then released, which showed President Bongo in good health in his address. But the 'oddness' of the video created doubt, and the military cited that oddness as evidence that President Bongo was not well and launched an unsuccessful coup¹⁰.

During the Brexit referendum social media platforms allowed each campaign to segment their audience into groups that were interested in different, specific, issues. Each group could be talked to in-

dividually, and exclusively of other groups¹¹. If you cared about animal rights, then you could be served adverts about how Brexit might advance animal welfare. Soon deepfakes will allow for those adverts to be AI-generated messages from politicians, or other recognisable figures, adverts designed to target increasingly smaller and more specific audiences.

The AI arms race

As much as AI can be used to create and spread disinformation, it can also be used to fight it. The latter is, however, much more difficult.

The first target of anti-disinformation campaigns might be what one considers to be 'inauthentic activity', such as spam posts by troll farm accounts. While some social networks like Facebook only want authentic users who represent real people, platforms like Twitter do not share this expectation. On Twitter, there is no preference for users to use their real names; a fact highlighted by the many satirical accounts which spoof real people.

The numerous legal and productive applications of deepfakes, from mobile apps like Snapchat to blockbuster movies, make it unreasonable to suggest that their creation ought to ever be illegal.

The ability of a country to quickly and thoroughly fight disinformation is a metric that will soon be used to know how likely it will be to keep stability. While the UK currently

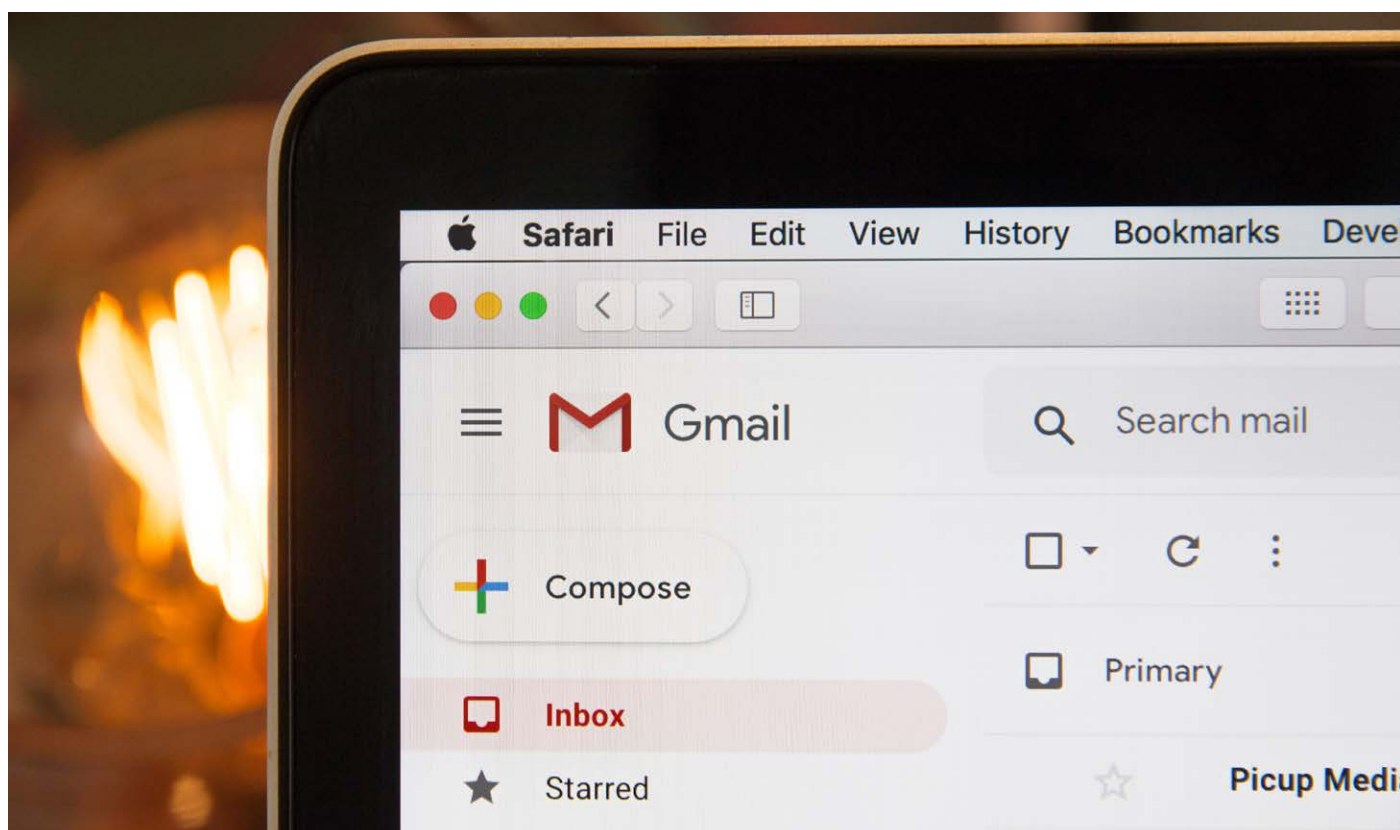
To remove or limit these accounts would be a direct blow against what the platform's users enjoy about them.

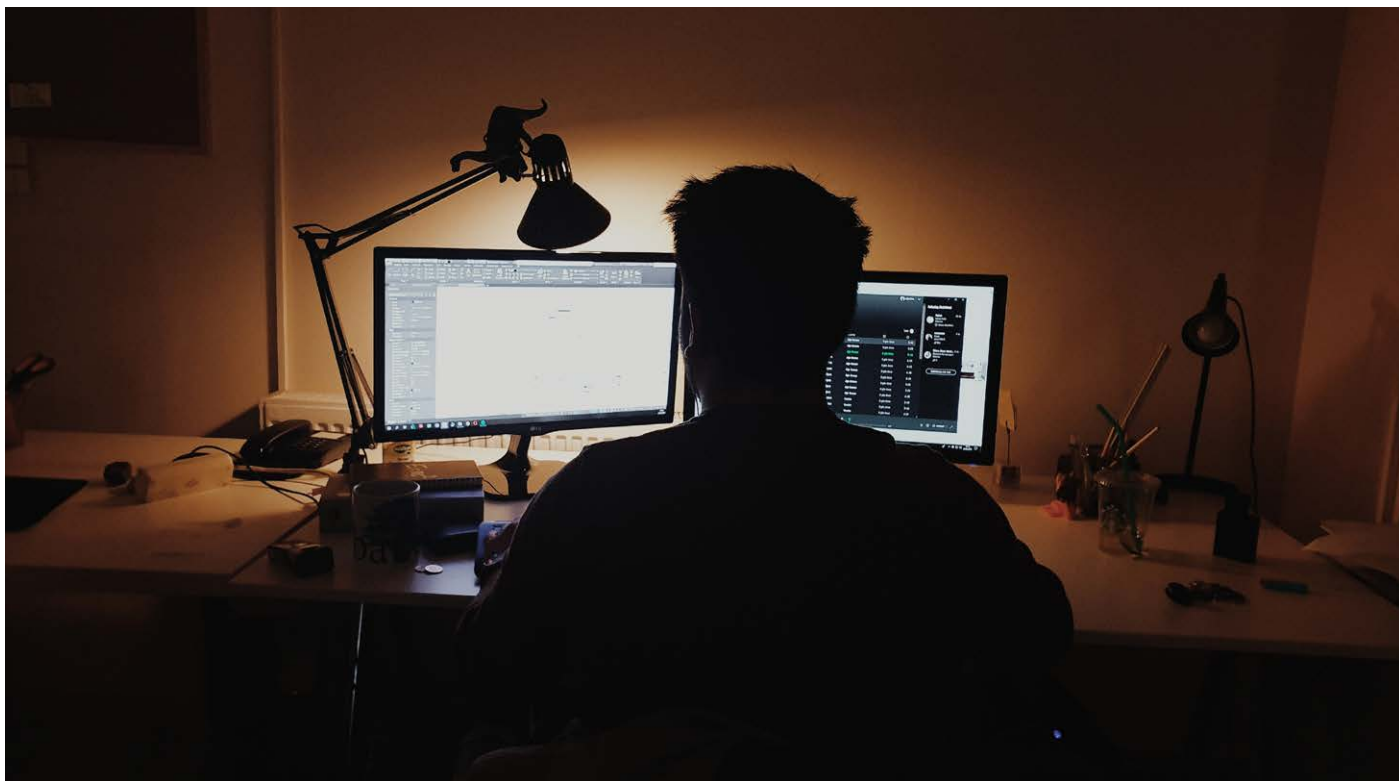
Instead of restricting satirical and 'anonymous' accounts, Twitter must look for behavioural patterns¹². These can point to coordinated inauthentic activity and are often signs of disinformation campaigns. AI can help to identify word patterns that can be indicators of disinformation and bot networks¹³. However, as a technology, AI pattern recognition is still developing. As such, it does not currently provide a complete solution to detect disinformation. It also still relies on

has the capacity to fight disinformation campaigns, deepfakes will allow for far more realistic and convincing fakes videos that will need more capable infrastructure to fight it. Additionally, the ability to fight disinformation in an ethical manner whilst preserving certain civil liberties will be a key test for liberal democracies such as the UK. 🇬🇧

human users to identify disinformation for training data and to make more complex decisions on the content that is flagged by this technology to avoid false positives. This human labour is often extremely manual, repetitive, and outsourced to developing countries

Algorithms are already being used to detect different types of content. Email spam filters, for example, are incredibly efficient at detecting spam emails. Advancements in the necessary technology will enable such detection tools to thrive further. 🇬🇧





Human moderation

Moderation of explicit content is difficult. For example, the line between art and pornography as established in the United States Supreme Court by Justice Potter Stewart was simply: 'I know it when I see it'.¹⁴

In many cases, human moderation is still used to identify content that is in breach of social media guidelines, including disinformation.

Human moderation has drawbacks, including the psychological cost. Moderators are frequently exposed

to images of a graphic nature, and the most common type of imagery used in disinformation campaigns is hate speech. Mueller's 2016 report, for instance, highlighted how race and gender were often used as ways to explore divisive issues in American contemporary politics. The psychological impact of constant exposure to material of this nature is demonstrated by the fact that Facebook moderators who have been exposed to this type of graphic and hate-fuelled content are now suing the platform as a re-

sult of developing Type 2 PTSD¹⁵.

Humans being exposed to disinformation, even as moderators, are still subject to the illusory truth effect and may come to believe the content they are being exposed to. In an investigation for The Verge, Casey Newton found that some moderators at Facebook had started to believe that the world was flat and 9/11 was not a terrorist attack, as well as denying the Holocaust¹⁶. 🚩

AI is not without its weaknesses

There are significant drawbacks with using AI to fight disinformation, including the underlying issues associated with the automatic moderation of free speech.

Any attempt at using AI ought to err on the side of caution and not be overzealous, as dealing with disinformation online is a persistent problem that cannot be solved, only stemmed.

False positives from AI tools are a threat to platforms themselves.

Over-moderation is antithetical to sites that thrive on user-generated content.

Perspective is the machine learning algorithm that Alphabet and Google use. Unlike many of Google's products, it is not an open tool. There are fears that tools used to moderate speech online – such as Perspective – can be misused or biased¹⁷. For example, it could be used by authoritarians to control speech, or a malicious actor to discriminate against a particular

minority. If Google was more open with how the tool worked, then the tool itself could be manipulated to be more prone to flagging the wrong kind of speech online. By giving the tool the wrong input data, it could easily flag dissent instead of disinformation. And because disinformation affects the general public, the status quo means the general public is putting their trust into Google. 🚩

Policy can forge a better future

There have already been serious impacts from disinformation campaigns. The coronavirus pandemic is an example of how any situation can be used as narrative fuel for a disinformation campaign.

Early on in the pandemic, conspiracies started to spread that 5G towers were in some way responsible for either the origin or spread of coronavirus. These conspiracies escalated to the extent that 5G towers were burned down by individuals who believed they were taking measures to protect themselves¹⁸. There is an increasing scope of damage that these disinformation campaigns could do in the foreseeable future through the use of AI. It is imperative that the government takes action through policy to contain the impact of these campaigns, and that it considers the benefits of using AI to fight disinformation.

With this in mind, policymakers must be careful not to stifle innovation in AI or bots. AI is a powerful tool to disseminate information and can be used for the public good as

an ‘early warning system for computational propaganda’ to stop disinformation campaigns before they go viral¹⁹. Many online bots can also be useful to facilitate this and are not weaponised to spread disinformation.

The most popular bots are often satirical and humorous. Dylan Wenzlau, founder of meme website *Imgflip*, used a natural learning processing algorithm to generate completely artificial memes that become popular online²⁰. Other playful bots provide services like tweeting emoji aquariums²¹ or randomly generating soft landscapes or star fields²².

There are more serious uses, too, such as bots that tweet whenever a Wikipedia edit is made from a New York Police Department IP address²³ or from the Houses of Parliament²⁴. These provide a level of accountability. It is not hard to imagine that journalists or investigators benefit from services that give them real-time updates of open source information as it happens or becomes available. It would not

be prudent to suggest that these bots, or any bot that does not represent a real or authentic person, should in some way fall foul of the law. Some bots must be protected.

But AI systems do not work independently of people, and a balance must be struck with the role of humans in AI. From selecting and generating training data to assessing the work that an AI has generated, humans have a role to play.

Policymakers must continue to ensure that humans double-check the results that AI produces in moderating free speech online. This must be balanced against the human and psychological cost that moderators face in being exposed to disturbing and misleading content.

If disinformation is seen as being truly opposed to the root of democracy, then fighting it can be viewed as a patriotic duty that must be supported²⁵. 🇺🇸

Bibliography

1. European Commission, A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation (Luxembourg: Publications Office of the European Union, 2018)
2. Special Counsel Robert S. Mueller, III et al., Report on the Investigation into Russian Interference in the 2016 Presidential Election, March 2019.
3. Christian Davies, 'Undercover Reporter Reveals Life in a Polish Troll Farm', The Guardian, 1 November 2019, <<https://www.theguardian.com/world/2019/nov/01/undercover-reporter-reveals-life-in-a-polish-troll-farm>> accessed 3 May 2020
4. John Seymour and Philip Tully, 'Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter', <<https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter.pdf>>, accessed 27 April 2020.
5. George Dvorsky, 'Hackers Have Already Started to Weaponize Artificial Intelligence', Gizmodo, <<https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425>>, accessed 27 April 2020.
6. Lynn Hasher, David Goldstein and Thomas Toppino, 'Frequency and the Conference of Referential Validity', Journal of Verbal Learning and Verbal Behavior (Vol. 16, 1977), pp. 107–12.
7. Robert Chesney and Danielle K. Citron, 'Disinformation on Steroids', Council on Foreign Relations, 16 October 2018 <<https://www.cfr.org/report/deep-fake-disinformation-steroids>>, accessed 3 May 2020.
8. Sara Fischer, 'How reporters outsmart the internet trolls', Axios, 17 September 2019, <<https://www.axios.com/reporters-trolls-news-media-misinformation-ae4a6e2a-1266-49bd-838e-d519588c-66cf.html>>, accessed 9 May 2020.
9. Giorgio Patrini, 'Mapping the Deepfake Landscape', Deeptrace Labs, 7 October 2019, <<https://deeptracelabs.com/mapping-the-deepfake-landscape/>>, accessed 18 May 2020.
10. Ali Breland, 'The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink', Mother Jones, 15 March 2019, <<https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/>> accessed 3 May 2020
11. Peter Pomerantsev, 'This is Not Propaganda: Adventures in the War Against Reality', Faber and Faber, London, 30 July 2019.
12. Evelyn Douek, 'Senate Hearing on Social Media and Foreign Influence Operations: Progress, But There's A Long Way to Go', Lawfare, 6 September 2018, <<https://www.lawfareblog.com/senate-hearing-social-media-and-foreign-influence-operations-progress-theres-long-way-go>>, accessed 27 April 2020.
13. Louk Faesen et al., 'Understanding the Strategic and Technical Significance of Technology for Security Implications of AI and Machine Learning for Cybersecurity', The Hague Centre for Strategic Studies (HCSS) and The Hague Security Delta, 28 August 2019.
14. Jacobellis v. Ohio, 378 U.S. 184 (1964), <<http://cdn.loc.gov/service/ll/usrep/usrep378/usrep378184/usrep378184.pdf>>, accessed 27 April 2020.
15. David Gilbert, 'Bestiality, Stabbings and Child Porn: Why Facebook Moderators Are Suing the Company for Trauma', Vice, 3 December 2019, <https://www.vice.com/en_uk/article/a35xk5/bestiality-stabbings-and-child-porn-why-facebook-moderators-are-suing-the-company-for-trauma>, accessed 27 April 2020.
16. Casey Newton, 'The Trauma Floor', The Verge, 25 February 2019, <<https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>>, accessed 27 April 2020.
17. Emily Dreyfuss, 'Hacking Online Hate Means Talking to the Humans Behind It', Wired, 8 June 2017, <<https://www.wired.com/2017/06/hacking-online-hate-means-talking-humans-behind/>>, accessed 27 April 2020.
18. BBC News, 'Mast Fire Probe Amid 5G Coronavirus Claims', 4 April 2020, <<https://www.bbc.co.uk/news/uk-england-52164358>>, accessed 3 May 2020.
19. Samuel Woolley, The Reality Game: How the Next Wave of Technology Will Break the Truth, (New York, NY: PublicAffairs Books, 2020).
20. This Meme Does Not Exist, <<https://imgflip.com/ai-meme>>, accessed 3 May 2020.
21. Emoji Aquarium, <<https://twitter.com/EmojiAquarium>>, accessed 3 May 2020.
22. Joseph Brogan, 'Some of the Best Art on Twitter Comes from these Strange Little Bots', Ars Technica, 6 July 2017, <<https://arstechnica.com/information-technology/2017/06/the-art-bots-that-make-twitter-worth-looking-at-again/>>, accessed 3 May 2020.
23. NYPD Edits, <<https://twitter.com/nypdedits>>, accessed 3 May 2020.
24. Parliament WikiEdits, <<https://twitter.com/parliamentedits>>, accessed 3 May 2020.
25. Cailin O'Connor, 'The Information Arms Race Can't Be Won, But We Have to Keep Fighting', Aeon, 12 June 2019, <<https://aeon.co/ideas/the-information-arms-race-cant-be-won-but-we-have-to-keep-fighting>>, accessed 27 April 2020.



Tom Ascott is the digital communications manager at the Royal United Services Institute and the associate editor of Intellect's journal Technoetic Arts. He wrote his graduate thesis on the risks posed by automated weapon systems. You can read more of his analysis on technology, disinformation and digital media on his blog: <http://www.tomascott.co.uk/>



SECTION 2

GLOBAL COMPETITION

THE ARTIFICIAL INTELLIGENCE LANDSCAPE

By Lewis Hammond

Introduction

AI has never been a greater focus of attention of the media, companies, and the general population than it is right now. The field has achieved remarkable successes in its short history, especially in the last decade, but from recent headlines one could be forgiven for thinking that we are on the verge of achieving Artificial General Intelligence (AGI). In order to cut through the ‘hype’ and to understand how and why the field is important, it is instructive to

consider its development and the current landscape of AI capabilities and research.

Here we provide a brief look at the history of AI, where it has brought us, and where we are likely to be heading in the near future. This assessment of the current landscape of the field serves as a background for the other sections of the pamphlet, including our policy recommendations. Note that the following (sub)sections are standalone, and so for the time-constrained

reader we recommend the following reading order (in increasing level of detail):

- Summary
- Overview (within Where Are We Now?)
- Where Are We Now? (which includes specific areas such as Academia and Industry)
- The Artificial Intelligence Landscape




History

Though related ideas had been considered since antiquity, the field of AI as a research programme began in 1956 at the Dartmouth Workshop in the US, attended by a handful of the world’s leading mathematicians, scientists, and engineers. Early progress was fuelled by intense optimism, large amounts of funding from government agencies in the US and the UK, and several successful applications in simple domains that required logical or symbolic reasoning such as simple algebra problems. In 1967 Marvin Minsky, one of the pioneers of the field, famously opined that “[w]ithin a generation ... the problem of creating ‘artificial intelligence’ will substantially be solved”. Needless to say, these expectations were far too high and in 1973 substantial funding cuts and criticisms were made that led to the first ‘AI winter’

– a period marked by comparably little progress, interest, or investment.

In the early 1980s Japan’s ‘Fifth Generation Computer Systems’ project sparked a new wave of funding from several other countries, including the UK. This, along with the development of expert systems and a revival of ‘connectionism’ (approaches based on neural networks), led to another period of intense interest in AI from government and (this time) business. Expectations were once again too high, however, and severe cuts to investment were made towards the end of the decade leading to a second AI winter. Despite this, steady progress was made by the field, marked by milestones such as the defeat of world chess champion Garry Kasparov in 1997 by IBM’s Deep Blue², and new approaches based on statistical methods and

intelligent agents that have paved the way for today’s AI systems.

At the beginning of the last decade developments in optimisation algorithms, powerful new hardware, and the availability of large, high-quality datasets combined to produce what some have referred to as the ‘deep learning revolution’. Many of the notable latest successes in computer vision, natural language processing, and game-playing (such as victory by DeepMind’s AlphaGo³ in 2016 against Lee Sedol, one of the world’s top Go players) are due to deep learning, a particular kind of machine learning, and the area has seen large amounts of research activity and investment in recent years. It is in many ways the driving force behind the current wave of AI and, as of the start of this new decade, shows few signs of slowing down just yet. 

Where Are We Now?

Overview

Cutting-edge AI development capabilities are currently concentrated in a relatively small set of companies and university departments, both of which have superior access to top research talent. The former also possesses the advantage of enormous quantities of data and amounts of computing power which are critical for many modern machine learning applications. These companies and universities are in turn concentrated in a relatively small number of countries. Particularly in the US, China, the UK, and Germany there is a significant degree of collaboration between academia and industry, with many top academics holding positions both in university departments and large tech companies. Excluding academia, corporate-affiliated AI research is more common in the US whereas government-affiliated institutions contribute the highest number of AI publications in China and Europe.

Both historically and presently North America is undoubtedly the

Academia

AI has grown significantly within academia in recent years with AI publications now making up 3% of all peer-reviewed journal publications and 9% of all published conference papers. A large portion of this growth is due to China, which now publishes as many AI journal and conference papers per year as Europe, having surpassed the US in 2006. The UK is well-represented at the international level, publishing the fourth greatest volume of journal papers

world leader in AI with the majority of the world's foremost AI companies (including Google, Facebook, Amazon, IBM, and Microsoft) and universities (including Stanford, MIT, Carnegie Mellon, and UC Berkeley). With that said, China now leads in the sheer quantity of AI research produced (though their increase in quality has been less significant, with publications from the US still being cited 40% more than the global average) and has been explicit about its desire to become the world-leader in AI by 2030⁶. Several large Chinese tech companies (including Baidu, Tencent, Alibaba, JD.com, and Huawei) alongside state-backed research programmes and data collection efforts have been critical to this impressive progress.


Within Europe, the UK leads both in the number of companies and of programmes offered by universities, hosting one third of AI companies and more than half of AI university programmes. It is also home to several of the world's top universities for AI (including Oxford, Cambridge, UCL, Imperial, and Edinburgh), influential organ-

and the sixth greatest volume of journal papers between 2015 and 2018. AI conferences (a strong indicator of both industry and academic enthusiasm) have also grown in size and prestige in recent years with attendance at the largest AI conference in 2019 (NeurIPS) up 41% from 2018 and up 800% from 2012 to approximately 13,500.

An important characteristic of AI in academia is that many resources are available freely online and there is strong support for open

isations such as the Alan Turing Institute (ATI), and arguably the world's leading AI lab: Google's DeepMind.

One of the driving forces behind the UK's AI ecosystem is the 2018 AI Sector Deal. This £1 billion package from government and industry seeks to address AI and data as one of four 'Grand Challenges' set by the Industrial Strategy white paper⁷, and has already contributed along each of its five target axes: ideas (new research on AI in healthcare, industry, engineering, and more), people (16 new Doctoral Training Centres, ATI fellowships, and industry-sponsored degrees), infrastructure (partnerships with the Open Data Institute and Innovate UK on three 'Data Trusts' pilots), business environment (establishment of the AI council), and places (publishing the 'AI Guide for Government'⁸ and establishing several new centres of excellence in data and AI)⁹. Though this is a positive first step, the growth of this ecosystem will require continuing support and regulation, and brings many new governance challenges that are yet to be addressed¹⁰. 

research. For example, many AI researchers now upload preprints online before they are accepted for publication, meaning that research is both more accessible and more quickly disseminated. Similarly, much research software and data is published online and there are large numbers of high-quality online courses in AI available at either zero or minimal cost, leading to an increase in self-study. 

The international facts in figures collated in this section are, unless otherwise indicated, taken from the AI Index 2019 Annual Report⁴, issued by the Stanford Institute for Human-Centered AI, which (as of writing) is the most comprehensive and up-to-date yearly survey of the global AI landscape. Further UK-specific information was gathered from the 2018 AI Sector Deal policy paper⁵. The interested reader is referred to these reports for further detail.

Industry

The growth in academia has been paralleled in industry, which is now by far the largest consumer of AI talent. In 2018, over 60% of AI PhD graduates went to industry, up from 20% in 2004. It is worth noting that an important consequence of the growth in the AI industry has been a significant ‘brain drain’ on universities, with large numbers of academic staff moving to private companies. This is hardly surprising when one considers the financial incentives: between January 2018 and October 2019 approximately \$2.9 billion was privately invested

Capabilities

There have been a range of recent breakthroughs whereby some AI systems now perform at or above human level on some narrow (though important) tasks. Examples include image classification (on several large datasets)¹², skin cancer classification¹³, poker¹⁴, Chinese-to-English translation of news stories¹⁵, optical

Ethics & Society

Public interest in AI has soared over the last five years and a similar picture can be seen in parliament, with almost 300 mentions of AI or machine learning in 2018 compared to fewer than ten in 2015¹⁸. Much of this interest, including this pamphlet, relates to the ethical use of AI. There have been at least 84 official proposals for ethical AI principles in the last few years alone from academic institutions, governments, indus-

try, and others, though concrete policies and procedures are still lacking¹⁹. Major concerns voiced in the majority of these proposals include Fairness, Accountability, and Transparency (FAT), as well as interpretability, data privacy, and robustness. However, such concerns seem not to be reflected in many of the companies using (but not necessarily developing) AI, with only 19% of businesses in the aforementioned McKinsey & Com-

pany survey saying that they are taking steps to mitigate risks associated with explainability of their algorithms, and only 13% mitigating risks to equity and fairness, such as algorithmic bias and discrimination²⁰. Government policy and regulation has so far failed to adequately address many issues in the ethical use and development of AI, despite their increasing importance. 🚩

in UK AI companies, behind only China (\$25 billion) and the US (\$36 billion). In 2019 alone private investment in AI globally was over \$70 billion, with AI-related startup investments over \$37 billion. This latter figure has increased at an average annual growth rate of over 48% since 2010 and continues to do so. Naturally, AI labour demand is also growing in significance, especially in hi-tech services and the manufacturing sector, and currently outstrips supply. From an international perspective, Singapore, Brazil, Australia, Canada and India

character recognition (on several large datasets)¹⁶, and many complex video games that are played professionally (which serve as a common benchmark in modern reinforcement learning)¹⁷. Other capabilities such as speech recognition are also nearing human levels in specific domains, however there are still many lower level or

experienced the fastest growth in AI hiring from 2015 to 2019. In a large-scale international survey of a wide range of companies (across multiple industries) by McKinsey & Company in 2019, 58% of respondents reported that they were using some form of AI for at least one function or business unit, increasing from 47% the previous year¹¹. This figure is similar across world regions, however different AI capabilities (for example, natural language processing or robotics) have been adopted to differing extents depending on the needs of the industries that are more or less

general tasks where AI performs poorly. This is commonly known as Moravec’s Paradox: the idea that the structured symbolic tasks (such as mathematical reasoning) that humans find difficult are far simpler to automate than the low-level sensorimotor tasks (such as catching a ball) that humans find easy. 🚩

general tasks where AI performs poorly. This is commonly known as Moravec’s Paradox: the idea that the structured symbolic tasks (such as mathematical reasoning) that humans find difficult are far simpler to automate than the low-level sensorimotor tasks (such as catching a ball) that humans find easy. 🚩

Looking Forward

It is notoriously difficult to predict progress in AI, but nonetheless there are certain technologies on the horizon that look set to have a large impact within the next five to ten years. Perhaps the most obvious of these is the deployment of autonomous vehicles (AVs). The

UK company FiveAI secured an extra \$41 million in funding earlier this year and have cemented their place as Europe’s leading AV startup²¹. Their first passenger trials on UK roads took place towards the end of 2019 with more due this year²².

A second AI-enabled technology that is likely to become particularly prevalent is sophisticated personal assistants. These are already present to an extent in forms such as Apple’s ‘Siri’ or Amazon’s ‘Alexa’, but continuing progress in natural language processing and an in-

crease in the number of compatible devices and services means that these assistants will become far more capable in the near future²³. Outside of these two developments we can expect to see continuing integration of AI systems in travel, healthcare, finance, and manufacturing, amongst other key industries.

With that said, there are still several important limitations to even the most powerful AI systems we have today, reflected by several open problems in the field. Modern deep learning models are data-hungry, difficult to understand, and lacking in robustness. Particularly in safety-critical areas, traditional techniques for verifying correctness do not yet scale to modern AI systems. These repre-

sent key challenges the research community faces in the coming years, although progress is already being made. While further off, the next frontier in AI systems arguably centres around common-sense reasoning, causal understanding, and the ability to generalise from little experience to new domains²⁴. Needless to say, progress on these more human-like abilities remains somewhat slower.

Finally, one might ask whether the overall rate of academic progress and industry investment is due to continue, or whether we face an imminent AI winter. The answer, for now, seems to point towards cautious optimism. AI capabilities are developing at a steady rate, and the lag of wide-spread deployment behind the cutting edge (which is

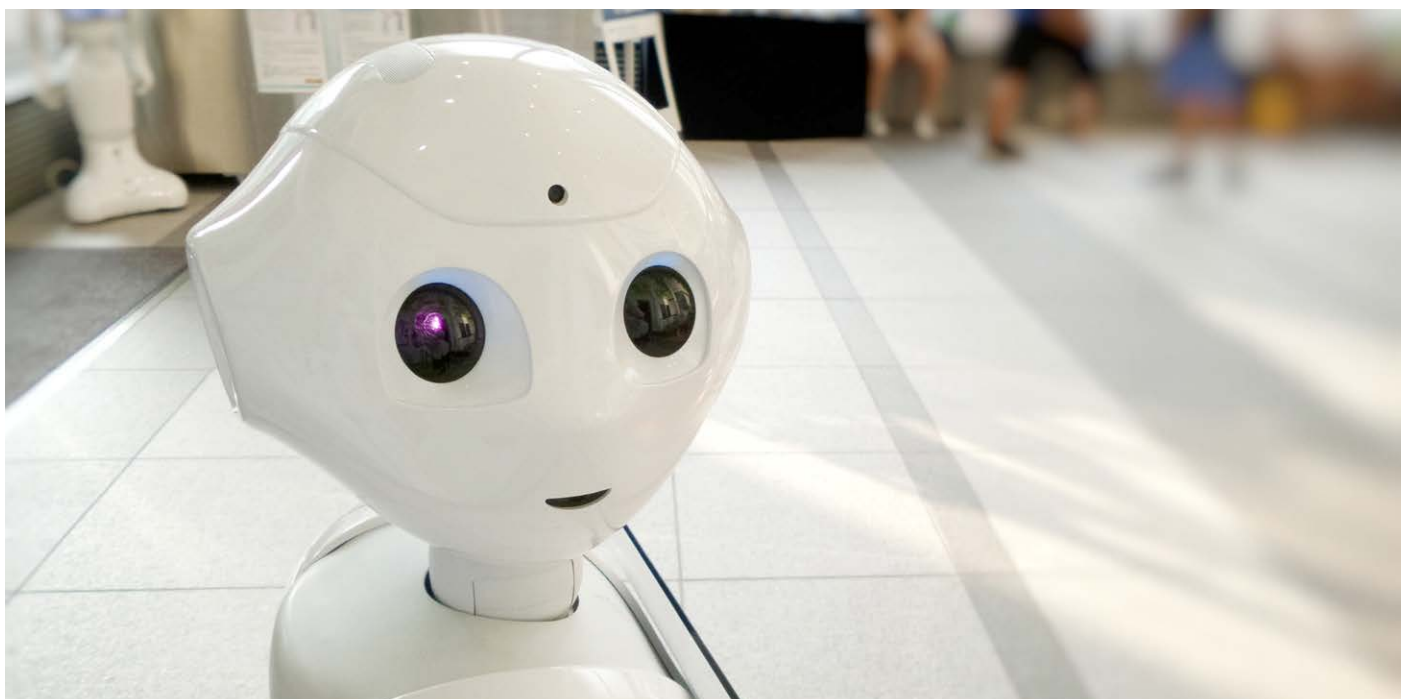
common in the tech sector) means that some of these developments are yet to permeate many areas. With that said, the history of AI teaches us that hype is not helpful, and that AI is not a panacea. There are also reasons to suspect that progress, while continuing, is slowing somewhat. Though the amount of compute used to train the most powerful AI models has been increasing exponentially²⁵, and algorithmic and hardware efficiencies have meant that training time is decreasing exponentially²⁶, it is currently unclear whether these exponential trends are reflected in the actual performance of the models. Assessing and predicting AI progress remains a challenging but important task in shaping its development. 🚩

Summary

Despite several slow periods, or 'AI winters', the field of AI has progressed rapidly in its short history and arguably never more so than in the last ten years. While previous disappointments caution against high expectations, there is reason to be optimistic that current trends will continue into the near future. Globally, the US remains the leader in both academia and in-

dustry though China has been successfully pursuing an agenda that hopes to see it overtake the US by 2030. In Europe, the UK is the top destination for AI talent due its universities and AI companies, both of which have been supported by the AI Sector Deal and will continue to be so for several years to come. There has been much discussion in public, Parliament, industry, and

academia about the ethical use of AI and potential challenges that the adoption of new, related technologies may bring. As of writing, it is fair to say that this discussion is only now finding its way into concrete policy proposals and regulation, and that there remains much work to be done in order to ensure that everyone benefits from the use of AI. 🚩



References/Bibliography

1. Marvin Minsky, *Computation: Finite and Infinite Machines* (Englewood Cliffs, NJ: Prentice-Hall, 1967)
2. "IBM 100 - Deep Blue", IBM, Accessed: 06.06.20, <https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/>
3. "AlphaGo", DeepMind, Accessed: 06.06.20, <https://deepmind.com/research/case-studies/alphago-the-story-so-far>
4. Raymond Perrault, Yoav Shoham, Erik Brynjolfs-son, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles, "The AI Index 2019 Annual Report", AI Index Steering Committee, Human-Centered AI Institute, Stanford University, 2019.
5. "AI Sector Deal", Department for Business, Energy & Industrial Strategy and Department for Digital, Culture, Media & Sport, 2018.
6. Jeffrey Ding, "Deciphering China's AI Dream", Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2018.
7. "Industry Strategy – Building A Britain Fit For The Future", HM Government, 2017.
8. "A Guide to Using Artificial Intelligence In The Public Sector", Government Digital Service and Office for Artificial Intelligence, 2019.
9. "AI Sector Deal – One Year On", Office for Artificial Intelligence, 2019.
10. Allan Dafoe, "AI Governance: A Research Agenda", Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2018.
11. "Global AI Survey: AI Proves Its Worth, But Few Scale Impact", McKinsey & Company, 2019.
12. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision* 115, 211–252, 2015.
13. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau and Sebastian Thrun, "Dermatologist-level Classification Of Skin Cancer With Deep Neural Networks", *Nature* 542, 115–118, 2017.
14. Noam Brown and Tuomas Sandholm, "Superhuman AI for Multiplayer Poker", *Science* 365, no. 6456, 885–90, 2019.
15. Allinson Linn, "Microsoft Reaches A Historic Milestone, Using AI To Match Human Performance In Translating News From Chinese To English", Microsoft, 2018, Accessed: 06.06.20, <https://blogs.microsoft.com/ai/chinese-to-english-translator-milestone/>
16. Claudiu Dan Ciresan, Ueli Meier, Luca Maria Gambardella, and Juergen Schmidhuber, "Deep Big Simple Neural Nets Excel On Handwritten Digit Recognition", *Neural Computation* 22, no. 12, 3207–20, 2010.
17. Dan Garisto, "Google AI Beats Top Human Players At Strategy Game StarCraft II", *Nature News*, 2019, Accessed: 06.06.20, <https://www.nature.com/articles/d41586-019-03298-6>
18. Raymond Perrault, Yoav Shoham, Erik Brynjolfs-son, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles, "The AI Index 2019 Annual Report", AI Index Steering Committee, Human-Centered AI Institute, Stanford University, 2019.
19. Anna Jobin, Marcello Lenca, and Effy Vayena, "The Global Landscape Of AI Ethics Guidelines", *Nature Machine Intelligence* 1, 389–399, 2019.
20. "Global AI Survey: AI Proves Its Worth, But Few Scale Impact", McKinsey & Company, 2019.
21. Matthew Field, "FiveAI Raises \$41m To Become Britain's Best Funded Driverless Car Start-up", *The Telegraph*, 2020, Accessed: 06.06.20, <https://www.telegraph.co.uk/technology/2020/03/04/five-ai-raises-41m-become-britains-best-funded-driverless-car/>
22. Mark Bridge, "Self-driving Cars Take To London Roads", *The Times*, 2019, Accessed: 06.06.20, <https://www.thetimes.co.uk/article/self-driving-cars-take-to-london-roads-ccx8ssksd>
23. Chris Arkenberg, "The Future of Intelligent Assistants", *Computer* 50, no. 12, 77, 2017.
24. Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. "Building Machines That Learn and Think like People", *Behavioral and Brain Sciences* 40, e253, 2017.
25. Dario Amodei and Danny Hernandez, "AI and Compute", OpenAI, 2018, Accessed: 06.06.20, <https://openai.com/blog/ai-and-compute/>
26. Danny Hernandez and Tom Brown, "AI and Efficiency", OpenAI, 2020, Accessed 06.06.20, <https://openai.com/blog/ai-and-efficiency/>



Lewis Hammond researches AI at the University of Oxford, where he is a DPhil student at the Department of Computer Science and an affiliate at the Future of Humanity Institute.

@lrhammond

THE EUROPEAN UNION: INTENSIFYING COMPETITION

By Antoinette Hage

Can We Cope Without EU?

The European Union, the United States and China are considered the major regulatory players which big tech fear. If this government is so ambitious about applying tech and AI for the common good, what strategy will it take to position the UK as a world leader in innovation?

The European Union has produced some of the most wide reaching regulation on technology and tech related issues in recent years. The General Data Protection Regulation (GDPR) outlined personal data rights for European Union citizens, giving them power to manage where their personal data is used or held. While the implementation has been imperfect, the spirit of the GDPR puts the citizen, not the tech firm, at the centre of the debate on personal data. The European Commission's Margrethe Vestager leads on the challenge of regulating big tech through competition instruments or content regulation, and she is feared to the extent that backchannel threats are made to prevent her exerting her power. Claude Moraes, former Member of the European Parliament from the UK, is one of few non-US legislators to hold Facebook CEO Mark Zuckerberg to account in front of a parliamentary committee. President of the European Commission Ursula von der Leyen has made digital transformation one of only two big issues her college of commissioners will focus on – the other being the green revolution.

In parallel, but perhaps not in contrast, the UK has the Information Commissioner's Office (ICO), the Financial Conduct Authority (FCA), Ofcom, the DCMS, the Office for Artificial Intelligence, the Cen-

tre for Data Ethics and Innovation (CDEI) and the Regulatory Horizons Council who offer up different policy ideas on areas of AI and data regulation. For example, the ICO is working on a long term AI auditing framework and the FCA on understanding the interaction of AI in traditional financial markets. DCMS led on the online harms white paper and regulating online content. The CDEI publishes snapshots on the use of AI and data in different sectors and industries. Ofcom acknowledges its role in the tech and communications space but hasn't yet figured out where it fits in the puzzle of innovation and regulation. While the EU and the UK could both be guilty of a disjointed approach to tech regulation, there is also some common ground in policy approaches and areas of interest.

Taking stock of the coronavirus pandemic, it became clear that the UK government as well as the European Union were not equipped to deal with the emerging health and economic crises alone. The UK Government turned to firms who could improve or accelerate diagnostics, contribute to a contact tracing service or produce ventilators and started recognising that digital exclusion was no longer acceptable. People cannot work from home without broadband. They cannot keep themselves safe without up to date information through the internet. The new Nightingale hospitals could not run without 5G masts. So, given the years of prior tech-lash and the drive behind the UK government and the European Commission to be seen to be 'regulating big tech', the relationship between the state and big tech

has changed. It can be argued that there is a new recognition, or a humility, that the state cannot do its job without tech firms of all sizes, and a further recognition that it does not have the expertise to compete with tech firms at the same level. Whether the state should have the same expertise as tech giants is another discussion.

How is the UK going to promote innovation and provide a sound framework for tech firms to grow and invest in Britain? How will we be more attractive than the European Union?

First, the UK needs to pledge data interoperability and equivalent standards of privacy with the EU. A tech firm is going to produce a product for the masses rather than produce different products for smaller communities with bespoke needs. As EU members, the UK was bound by the same data and privacy standards but outside of the EU, divergence from these standards may make the UK less attractive than its continental neighbours. Tech firms can work with tough privacy standards, but what they cannot work with is ever changing standards or a lack of standards that result in junk data. If the UK is to attract innovators on par with the EU, data interoperability and equivalent privacy regulation with the EU is a gold standard.

Second, the UK needs to embrace a digital ecosystem as the primary ecosystem. The UK lags behind the European Union in this respect. Estonia developed a comprehensive and world leading digital citizenship system beginning decades ago, enabling any individual to connect bank accounts, benefits



and taxes under one central digital ID. Their embrace of the future produced a thriving private tech sector as well as efficient digital public services in a small country. The European Commission has touted the green revolution and the digital revolution as the two central pillars of their work up to 2024. Contrast that to the UK, where there are whispers about embracing a digital ecosystem as a primary ecosystem, but where the political and policy will has lacked. The disjointed thinking across departments and regulators creates confusion and overlap. The talk about leveling up the regions sounds as if it is going to happen outside of a digital ecosystem, and yet UK banks are not being held responsible for the excruciatingly slow rate of approving Bounce Back bailout loans needed to save the businesses that could level up the regions. This is a process that could have been simplified and digitised, and it hasn't been - and the smallest businesses in the toughest circum-

stances have suffered. Antiquated systems, public or private, can be seen as a result of previous decades of underinvestment and lack of focus on new technology. More recently, the EU's COVID-19 recovery plan has dedicated chapters about the use of a digital ecosystem in the recovery, especially for small businesses to unlock improved productivity. The UK Government's exit plan only seems to want to reverse the lockdown to end up where we were previously, and there is growing sentiment that that is not where we want to be in the future.

Whereas the UK previously worked within the EU on digital policy, AI regulation and data privacy, this is a new step standing alone on a global stage. The Government admits their limited tech expertise and continues to marginalise innovation and digital public services. If a tech firm is going to set up a home in the UK market, it needs to know that the state will support

its innovation, that it can grow productively and sustainably, and that it can freely access the European Union single market.

There cannot be a green revolution without a digital revolution. We cannot build a smart energy grid if we cannot build the networking infrastructure or use and develop the AI needed to run it. We cannot put driverless cars on the roads without a sandbox for testing their AI in a safe environment. We cannot responsibly innovate in healthtech and in our NHS without balancing privacy needs with data sharing and data adequacy protocols. The Labour Party has to mount an effective opposition with internal expertise on digital issues backed up by strong relations with tech firms. We need to look at hiring people who could get jobs in tech firms, using their expertise to make policy and to better understand the needs of tech firms and what would attract them to invest in the UK. 🇬🇧



Antoinette Hage works in political and policy affairs, advising on developments for tech, fintech and financial services clients in the UK. She holds an MA from King's College London in Political Economy. She has previously worked for Hillary Clinton, the International Planned Parenthood Federation, the Labour Party and Britain Stronger In Europe.

AMERICA: BIRTHPLACE OF THE INTERNET

By Hannah Fuchs

It is vital to recall that the Internet, radar, bluetooth, mobile phones, etc. were all US state-supported inventions. The US state helped cement global dominance in those emerging fields through funding research and building AI technologies. The Internet itself was funded by the Advanced Research Projects Agency (ARPA), part of the Department of Defense and nowadays referred to as DARPA. Created by the Eisenhower administration in 1958, it led to a new industrial revolution. Most of DARPA's spending is part of the Pentagon's infamous black budget. It would not even confirm programme code names "or confirm estimates of the agency's bottom line." Since 2005, DARPA has been more transparent. In 2019, its enacted budget was \$3.427 billion.

In 1956, the US computer scientist John McCarthy organised the Dartmouth Conference, where the term 'Artificial Intelligence' was first adopted. Since then, researchers have tried to build intelligent machines but have failed, partly due to a lack of required large amounts of data (computers up until the early 1990s didn't have the capacity to store and process the amount of data needed).

Over the past few decades, research and development of many AI technologies in the US have moved from state-funded projects to private companies that benefit from the positive feedback loop of having exclusive access to an outstanding amount of data. The US state doesn't regulate the use of that data to a large extent (and what regulation exists is splintered over hundreds of individual state and federal laws) and lacks in antitrust enforcement, making it easy for these conglomerates to simply buy out their smaller competitors.

It is nearly impossible for start-ups to compete alongside Amazon, Google or Facebook. Sooner or later, they will receive an inquiry about selling their company.

The lack of state intervention (i.e. the regulation to protect private citizens) and the power that US companies using AI have, demonstrates the ethical risk of a positive feedback loop: the more data companies gather, the more of a monopoly is created. They can then drive down prices and eliminate competition. This concentration of power will feed into social inequality. It is therefore important to strengthen a diverse AI industry where small start-ups have a chance to compete on the market.

The increase of 'gig economy' or 'gig workers' has led to a real issue in the US. Gig workers are not a strictly defined category so assessments of their number vary. Generally speaking, gig workers are freelance workers that more and more often work in the platform economy, such as for Uber, Lyft or food delivery companies. Because those companies offer a platform where supply and demand meet, they do not see themselves as conventional employers and shirk their responsibility to provide platform workers with health insurance, minimum wage and paid leave. They also often don't receive unemployment insurance from the government as their status (whether they're an employee or freelance workers) is ill-defined. Gig workers find themselves in a gray zone with very few rights. In response to that, the state of California passed a law at the end of 2019 that classifies workers in the platform economy as employees, enabling them to receive valuable workers' benefits. The traditional

employer-employee relationship starts to blur with artificial intelligence. When apps, or algorithms, make the decisions, then who is accountable for their impacts?

Policymakers need to understand the economic disruption that comes with AI.

- Rather than looking away, the most beneficial outcome for society is for national as well as regional governments to understand new economic opportunities and proactively create jobs that allow employees to work with AI rather than have AI replace jobs. Innovations often come along with economic disruptions. The government should not be facing mass unemployment, but instead, mass redeployment.
- Governments also need to ensure that people don't end up in gray zones with little rights. Instead, as mentioned earlier, minimum wage, health insurance, pension, as well as paid leave should all be guaranteed for so-called 'gig workers'. The definition of this group needs to be revised regularly due to the fast changes happening in this industry.
- The UK government needs to ensure that monopolies are properly identified and regulated so that a healthy competitive environment can exist alongside workers' rights. This can be done, for example, by institutionalising a wide range of easily-accessible and low-threshold funding opportunities for start-ups, and redistributing the resources of large tech

companies that already have access to an extraordinarily large pool of data. Those resources can then be used to retrain people, create new job sectors for those whose jobs are at risk, and invest in smaller start-ups.

In an ever more globalised world, there is a great need for international cooperation on taxing tech companies to reduce inequality on a national as well as an international level (e.g. avoiding tax havens).

Currently, the US remains the global leader in AI, especially in chip development. China is well aware of that dependence on US-developed chips and is eager to catch up in that field, fueled by the US-China trade war and accusations against Huawei, a Chinese telecommunications company, which allege that the company helps the Chinese government steal foreign AI technology and allows Chinese intelligence agencies to use their telecommunication networks to spy on foreign countries. A recent study indicates that the US is going to fall behind China in five to ten years in the areas of innovation, implementation, and investment of AI. In the same study, the UK follows China ranking third. Research on the US AI workforce conducted by the Center for Security and Emerging Technology states that more US companies are moving their AI research and development (R&D) abroad and 70% of computer scientists studying in the US were born abroad.

In light of Brexit, the UK government will need to consider its immigration policies and create an inclusive environment both for national and international students. It should encourage UK residents along with international students to enroll in STEM programmes to improve and encourage diversity in the industry and ensure that the sector reflects the makeup of the

general population.

As China catches up in AI research and development, the lead that the US has, is being slowly eroded. Experts and scientists have been criticizing this development, urging the US government to take action. In response, in 2019, President Donald Trump issued an executive order to make AI research and development a national priority. Part of that is a budget proposal intending to raise DARPA's spending on AI research and development to \$249 million from its current levels at \$50 million. The proposal also includes a budget increase for the National Science Foundation for AI matters. For many years, the US government has not only lost highly qualified data scientists and start-ups to China, but also to its private US-based companies such as Google. With the overall \$4.8 trillion heavy budget proposal, the US government aims to attract well-trained scientists and drive research in a direction that benefits national security and other public areas such as energy.

The US government struggles to keep on top of its US-based tech companies such as Google, Facebook, and Amazon. Most innovation has happened in the private sector and the US government does not seem to understand the consequences of AI in order to regulate it accordingly. At the moment, it looks like AI is regulating the government rather than vice versa. Governments need to understand AI and its impacts to the best extent possible in order to effectively promote and regulate it at the same time.

While the US government's R&D has lagged behind the private sector, its willingness to cooperate with it exists, raising questions on human and civil rights. For example, the technology company Palantir Technologies provides AI database management to Immigration and Customs Enforcement (ICE), part

of the Department of Homeland Security (DHS), in order to record, target, and detain undocumented immigrants. They were able to do so due to a cooperation with the FBI, where they used an unprecedented surveillance infrastructure that was based on facial recognition technology.

This raises the question of how far governments can use their power on their own citizens. For example, do a vast majority of undocumented immigrants represent a threat to national security? Or have undocumented immigrants actually contributed to their communities, and aimed to have a better life? Eventually, similar AI-based procedures will most likely be used on US residents as well, questioning the transparency, and legality of its usage. It bears the risk that people's lives will be heavily influenced by nontransparent algorithmic calculations rather than people's individual choices.

When such significant decisions about humans' lives are made, they shouldn't solely be based on algorithmic determinations. Instead, decisions should also be made in conjunction with ethical and moral stances that are transparent enough so that third parties, such as civil rights organisations, can ask questions and receive answers. Public institutions should also lay out on which premises their decisions were made to ensure transparency and allow the public to understand them and give the opportunity to question these decisions.

Eventually, consistent with the model in existing western liberal democracies, the US ought to move towards a model where the ethical boundaries surrounding the applications of AI are representative of the general public. Whether it can achieve that goal with its existing political system is another matter. ■

EAST ASIA: POWERFUL STATES

Singapore: The Lion City

Even though Singapore might not be at the top of the list when we think about AI, the city-state shows examples of how states can engage productively with the rise of AI. Similarly to the US and China, Singapore understood early on that AI could be a key driver for economic growth. In 2017, the Singaporean government set up a national programme to invest \$150 million into AI for the next five years. Its three key sectors are finance, city management and healthcare.

One year later, in 2018, an advisory council for the government was established, led by former Attorney General Vijaya Kumar Rajah. Its purpose is to advise the government on AI and work together with the ethics boards of businesses. Leaders from Google, Microsoft

and Alibaba are part of the advisory council. The potential economic gains from AI can conflict with the need for independent and universal ethical standards.

In November 2019, the government expanded the scope of its focus on AI and defined five key sectors: transport; smart cities; healthcare; education; and safety and security. Part of its national AI strategy is its Model AI Governance Framework whose second edition was launched in January 2020. The platform is aimed to democratise AI technologies and their use by implementing four principles:

1. Ensuring transparent internal governance structures and regular staff training within organizations
2. Determining the level of human involvement in AI-augmented decision-making so that organisations minimise the risk of harm to individuals
3. Minimising their biases in data and models
4. Communicating openly and accessibly with their stakeholders and allowing feedback

Even though it remains an open question as to how effectively this framework will be implemented, the principles should be adopted by all public and private organisations to ensure AI is used in a transparent and democratic way. In the midst of success due to AI, following ethical guidelines can often be overlooked by private companies as well as the government. 🇸🇬

China: from an agrarian state to the United States' biggest competitor in AI

Well into the latter part of the 20th century, China was still considered an agrarian state. Today, China is the world's largest producer of digital data, a gap that is widening daily. While China is well aware that in order to be able to compete in AI, it needs highly educated technical talent, it also knows that highly educated data scientists will reach a certain threshold whereby they begin to show diminishing returns. Beyond that point, data makes all the difference as it is the fundamental component without which AI could not exist in the first place.

According to Kai-Fu Lee, Founder, Chairman, and CEO of Sinovation Ventures, and the former president of Google China as well as executive at Microsoft, SGI, and Apple, China identified the four compo-

nents to be successful in AI: entrepreneurs, enormous amounts of data, highly educated AI engineers, and a government that is eager to support and use AI technology.

China has more internet users than the US and Europe combined. It has also started earlier in collecting a high quality of data which will be more useful for creating AI driven products. "Qualitative" data implies information collected from the real world, that is, physical purchases, meals, makeovers, transportation, etc. The higher the amount and the more wide-ranging the data, the better the data-fueled algorithms and models for future products will be.

The mobile app WeChat is one example of a successful product in China that is based on a data eco-

system. The app reflects the network effect stemming from its ecosystem. As WeChat affects almost all parts of Chinese life, the app constitutes an ecosystem. It is so tightly interwoven with day-to-day activities that it is challenging for a Chinese resident to not be part of this network. Additionally, because WeChat has a large amount of users sharing their data in so many areas of life, it can build on that network of information to further improve the app's services, or improve targeted advertising.

WeChat became a data powerhouse and, critics say, a tool for "remote control" of people's lives within just five years of launch as it enables messaging, media, marketing, gaming, payments at restaurants and your taxi driver, unlocking shared bikes, managing

By Hannah Fuchs

investments, booking GP appointments and having prescriptions delivered to your door.

During the Covid-19 lockdown in China, the State Council introduced an app based rating system together with two major tech firms, Alibaba and Tencent, to control people's movements during the outbreak. People log their recent locations and health statuses and the app is able to link all entries together. Based on that information, the users receive coloured badges (green, yellow, and red) and a QR code to show when, for example, they enter a building. Tencent and Alibaba appreciate the traffic on their systems but claim to not have any access to personal data. This use case is an example of one of the potential drawbacks of AI - opaque decision making. The Wall Street Journal reported on the case of a man who was granted a red badge despite following all instructions. Additionally, since the system's core operation is managed by the Chinese government, which means that the state now has (in China's case, virtually uninhibited) access to this powerful tool. For example, the provincial Hangzhou government accused 16 people of lying about their health conditions and immediately gave them red badges.

Because people use WeChat for every aspect of their lives, all their behaviours, patterns, and choices are recorded and centralised on the app. Every move, every decision, and even every thought that you type out and send to your friends, will be stored and used in an unknown way and by unknown people. That extreme centralisation of people's lives' data in one place is unique and creates a positive feedback loop; more services offered in one single place or app leads to more centralised data which leads to better products, which leads to more users, and so on.

The rise of WeChat, an app with an

enormous ecosystem, also spear-headed e-commerce in China. Targeting customers became effective and efficient because all their data was already provided through other applications within WeChat. China's digital transaction value in 2019 amounted to \$1,595,513 million, compared to \$152,897 million in the UK. China's mobile transaction penetration rate is higher than in any other country (35% vs. 6.6% in the US). That means 35% of people using a mobile phone also pay by using their phones - about half a billion people in China make their daily purchases by phone. Their average annual transaction value (\$1,662) is lower than that of the US (\$2,993) or UK (\$2,464). Although this figure is lower than those for the West, it is worth remembering that Chinese incomes are significantly lower than Western ones in absolute terms, making these figures all the more surprising. It reflects the increasing trend in China to pay all your daily purchases, no matter how small, with your mobile app. This development leaves behind an enormous amount of digital footprints of everyday behaviour that is stored in, centralised in, and thus made available to apps beholden to the Chinese state.

WeChat uses people's data from their everyday lives for extremely targeted advertising. How are WeChat users able to make critical, informed decisions if they get recommended products and services that seem plausible to them? And who designs those algorithms that offer you exactly that product that you allegedly have been looking for? The growing shift from contextual to behavioural targeting results in a continuous subtle influence of showing people certain products, services, news headlines, or bargains over and over, thereby influencing the way they perceive the world and make decisions. Global conglomerates can dominate the market by paying enough money and using their large sets of

available data as barriers to entry for startups. At some point, smaller companies most likely won't be able to compete anymore if no regulations cut off this cycle of data collection and competitive advantage entrenchment.


In 2014/15, the country became a real competitor to Silicon Valley with China's mass innovation campaign by focussing intensively on the following policy areas:

- The state started directly subsidizing technology entrepreneurs
- Public venture capital funding jumped tremendously and became almost equal to the US in 2018. Kai-Fu Lee points out that in America, people predominantly believe in private rather than public venture capital, as they tend to believe the latter is highly inefficient
- The establishment of entire cities focussing on AI. While the direction originated from the central government, ambitious mayors implemented the strategy widely. They aimed to establish their towns to be centers for AI by investing in local AI companies, offering research grants, opening AI training institutes, free company shuttles, securing places at schools and special accommodation for people who work in the AI industry
- The amount of technology incubators was rapidly increased. "Entrepreneurship zones" were created and government-backed funds were launched to attract more private venture capital. The government also granted tax incentives for people and businesses working in the technology sector and generally made it easier to start a business

Even though it sounds like a prom-

ising and successful strategy, how far can a government go in directing national industrial strategies? The above scenario bears the risk of a two-class society; those who

work in AI and those who don't. While these policies should be incorporated in the UK's AI policy strategy, policymakers will have to ask themselves how far they

can incentivise one area without groundlessly disadvantaging people at the same time. 

Online to offline revolution

The development outlined above, with the rise of WeChat, and what is part of the AI revolution is the introduction of the online to offline, or O2O Revolution; offline and online would merge together and there would be no more differentiation between what was online and what offline. The US introduced the first transformational O2O model: ride-sharing, thanks to Uber and Lyft. China quickly copied that model with Didi Chuxing and accustomed it to local conditions. WeChat then accelerated the O2O trend. An increasing amount of activities you do offline is managed online in one single app, offering all the services you need and thereby transforming the data environment.

As we know, WeChat centralised all its data gathering on consumption patterns and personal habits. That ecosystem differentiates China from the US, which doesn't centralise multiple services but instead, splits them up, offering multiple services across their different platforms. Facebook is an US example that splits its services into the Facebook app, Facebook Messenger, WhatsApp, and Instagram. Facebook even has its own app for managing pages and groups. All of these platforms seem to be independent and yet, all are owned by Facebook. On the other hand, Yelp bought Eat24, a food delivery platform, trying to follow the example of Chinese companies. However, it failed to properly fuse all of the logistical services onto one platform like Chinese companies do. Specifically, the restaurants still had to handle the deliveries themselves, which gave little incentive to join Eat24 and thus, the business never succeeded.

China was also able to catch up with the US so quickly because AI researchers around the world are relatively open to sharing their data, algorithms, and results with the public. Open source platforms (such as well-known Wikipedia) have become more and more popular. Publicly available knowledge across the world fosters competition on an international level. China put that to good use and proved to be a serious competitor in the field.

Another central component for AI is chip development (e.g. for facial recognition or self-driving cars). Even though Silicon Valley remains the clear leader in AI chip development, Chinese cities have become AI development hubs due to the following supportive policies:

- Easily accessible subsidies for research
- Venture capital funding and grants for AI companies
- Government contracts promising to buy products and services developed in local AI cities
- AI incubators
- AI training institutes
- Clear schemes to set up and register a company

The measures taken by the Chinese government raise questions about how independently firms can really operate. For example, an official statement laid out that government representatives would be assigned to 100 big tech companies including Alibaba in order to strengthen government relations and information exchange. It is not clear though to what extent the Chinese government is controlling these companies on the manage-

ment side. This is a democratic and transparency issue.

While the government does play a crucial role in helping start-ups to become successful, the public has a right to know by whom AI companies are funded and supported by. Knowing who controls the data collection and builds the algorithms is essential for ethical AI practice.

One particular categorisation of AI splits it into four. First, Internet AI uses data for algorithms to develop recommendations for users, such as seen with Youtube videos and Spotify songs. Second, companies use business AI to learn more about their customers to improve their services. For example, banks give out loans, insurance companies sell policies, or supply chains and inventories are getting optimised based on structured data that identifies certain patterns. Here, the US is the clear leader where companies specialise in helping other businesses improve their services through artificial intelligence software. China has so far been lagging behind here. Third, perception AI digitises the physical world, and how we perceive and experience it. It incorporates our daily routines, behaviour, and conversations by deep learning algorithms into data sets that can then be used in a wide variety of ways. Examples are Alexa, Siri, or the leading speech recognition company iFlyTek from China. Fourth and final, autonomous AI is slowly developing, such as self-driving cars, autonomous drones, and intelligent robots.

Ethics plays an especially crucial role in perception and autonomous AI. By gathering data in public spaces, questions arise such as how people give consent. Who can use

that data and for what purposes? How would checks and balances work? Among Western countries, the UK is already a leading country in public surveillance through the amount of closed-circuit television (CCTV) cameras in public spaces that feed the information into facial recognition software. The system has attracted heavy criticism over the years, including a 2018 judgement by the European Court of Human Rights ruling that the way data is collected has been unclear and therefore violates human rights. After Beijing, London is the city with the highest amount of CCTV cameras with around 420,000 cameras in 2019.

China is also advanced in autonomous drone production. The world's leading consumer drone maker, DJI, is based in Shenzhen and holds an estimated 70% global market share. The US has become sceptical about using their drones for government purposes due to security risks and DJI's alleged links to the Chinese government.

In general, Kai-Fu Lee argues that the US and China have different approaches to entering the market. The US is a "perfectionist", working on a product in Silicon Valley until it is nearly flawless, before it is rolled out around the world as an "one size fits all" product. China, on the other hand, uses a more diversified approach by investing in dispersed small local start-ups around the world, adapting the product's algorithms with local data, and tailoring it to local circumstances.

China has been successful in catching up in AI with an extreme pace, but that isn't to say that all developments in AI in China have

been good. For example, the Chinese national police use facial recognition technologies to target Uighurs, a minority group in China. In 2019, The New York Times reported that "Almost two dozen police departments in 16 different provinces and regions across China sought such technology beginning in 2018, according to procurement documents. Law enforcement from the central province of Shaanxi, for example, aimed to acquire a smart camera system last year that 'should support facial recognition to identify Uighur/non-Uighur attributes.'" While this is an example of discrimination, the Chinese start-up CloudWalk openly advertises that its surveillance system can "identify sensitive groups of people". As Clare Garvie, an associate at the Center on Privacy and Technology at the Georgetown University Law Center states, "If you make a technology that can classify people by an ethnicity, someone will use it to repress that ethnicity."

While China's approach in catching up with AI shows a holistic approach, and drives the development of AI on multiple levels and with incredible speed, the use of power China gained in AI remains rather questionable. Authoritarian states have an advantage in collecting data as they face less legal constraints. Gregory Allen, a political scientist and Chief of Strategy and Communications at the DoD Joint Artificial Intelligence Center, says "essentially all major technology firms in China cooperate extensively with China's military and state security services and are legally required to do so. Article 7 of China's National Intelligence Law gives the government legal au-

thority to compel such assistance, though the government also has powerful non-coercive tools to incentivize cooperation".

The UK will have to consider how to interact with China on AI matters:

- The private as well as public sector need to stay aware of AI developments in China, knowing that companies collaborate closely with the Chinese government. This means looking at the import and export of products and services to and from China, but also other states that knowingly have been supported by Chinese AI companies.
- The UK should have clear standards on human rights violations. While it is crucial to remain diplomatic relationships, the UK should openly speak up on human rights abuses.
- With the right set of policies that foster AI R&D, wealth distribution and the support of start-ups and small and medium sized enterprises (SMEs), the UK should cooperate with other nations and multilateral institutions to establish a level playing field that allows fair competition and protects human rights.
- By introducing the above point, the UK should aim to avoid a trade or proxy war with undemocratic states such as the US has been with China. 🇨🇳

AI IN FOREIGN POLICY



Artificial Intelligence is not only used for commercial purposes: the mobile phone, bluetooth, GPS, and the Internet are examples of inventions massively fund-

ed by and for the US military.

The uses of AI in foreign policy have been recognised and actively utilised by a few states for decades. Just like companies can

become monopolies in AI due to an extremely large pool of data, so too can world powers shift, vacuums be created and new tools of foreign policy established. 🇷🇺

Democracy, sovereignty and ethics

China has become a significant investor in American start-ups that are working on technologies with potential military applications. These start-ups focus, for example, on rocket engines for spacecraft, sensors for autono-

mous ships, and printers that make flexible screens that could be used in fighter-plane cockpits. Many of these Chinese investor companies are state owned or have connections to Chinese leaders. Not only does that mean Chinese investors

have significant control in the start-up's decision making, and deciding in which direction the start-up should go, it also opens the door for intelligence gathering and obtaining the technologies themselves. Especially when it comes to

military purposes, this can become critical.

Foreign direct investment (FDI) should not be underestimated as a foreign policy tool. In terms of the ethical use of AI, the public of one country (here the case of the US) often doesn't know (yet has a right to know) that the company they're giving their data to (here, China) is owned by the (Chinese) government.

- Because data can be used in such a wide variety of ways and the line of when AI becomes harmful is often blurry, it is extremely important to oversee who collects the data and how it is used.
- Policymakers and public servants are obliged to work in the public interest. If foreign states decide to interfere through FDI and have the ability to shift foreign policy tools without the oversight of domestic policymakers, then this raises questions on the ethical use of AI, and the undermining of public trust and democratic processes.

The UK's foremost AI company, DeepMind, has some of the world's leading AI scientists. In 2014, it was acquired by Google.

- The UK government should invest in and protect independent and UK-based AI companies in order to increase their global competitiveness.
- It should also remain in control of how the wealth generated through these companies is distributed throughout British society. After all, AI companies generate their wealth through taxpayers' data.
- Large AI companies need to be taxed to ensure a competitive market: first, to ensure enough is "paid" to users for their data that they are sup-

plying to AI companies, and second, to counteract monopolies created through the positive feedback loop introduced earlier in order to support a wide variety of AI start-ups. This payment ideally takes the form of market-determined negative prices or in the form of data to ensure a level playing field (see Conclusions: Policy Proposals).

Hacking foreign democracies poses another threat in foreign policy. Russia did so in the US elections in 2016 and in 2014 Chinese hackers stole files of 22 million people from the US government's Office of Personnel Management. China now could use this well-structured data to create algorithms in an extremely large variety of ways, thereby strengthening their cyberwarfare capabilities in many ways. In the 2016 US presidential election, bots were able to alter entire national public debates, and change people's opinions (see AI and Disinformation). In the 2020 US presidential election, presidential candidate Joe Biden has faced hacking attempts by Chinese hackers, targeting the personal emails of his campaign staff members. Bots can work 24/7 and process data as well as develop content in a much more efficient manner. The fact that an entire democratic system can be undermined by such attacks so easily shows how vulnerable societies are and how urgently states need to work on protecting the essential principle of societies living together: democracy.

This new form of AI attacking is called cognitive hacking, "a form of attack that seeks to manipulate people's perceptions and behavior, takes place on a diverse set of platforms, including social media and new forms of traditional news channels". Cognitive security, on the other side, aims to defend such attacks. Cognitive hacking abuses a large amount of innocent people and their data and engages them

in operations of foreign states without their knowledge.

In 2017, Google signed a contract with the US Department of Defense (DoD) for a military project called 'Project Maven', which deploys AI to "automatically label images, buildings, and other objects captured by cameras on drones, helping [US] Air Force analysts identify unique targets." It is an attempt to incorporate AI into battlefield technology. When the contract became publicly available, Google employees protested, and some quit their jobs while others started a petition to urge Google to distance itself from warfare technology and cancel the contract. At first, Google tried to play down the significance of the contract, saying it was "only" a \$9 million project. However, it was soon revealed that the contract with Project Maven was worth around \$250 million a year. In June 2018, Google announced it would let the contract for Project Maven expire when it ended in March 2019.

Employees should have the right to easily opt out of projects which go against their moral beliefs without consequential disadvantages for their career. This also raises the question of how strictly divided the lines between private companies and the government should be. Is Google allowed to share its users' data with the Pentagon without the users' consent? If so, how many third parties are allowed to access and use that data, and then use it for what purposes? Or should it not be possible for Google to share the data at all?

Google is also an international company with offices across the world, which entails two kinds of risks. First and foremost, Google collects data from countries all over the world. Despite GDPR rules, there is leeway for Google to use data from its users in foreign countries for US military purposes. The second risk is that classified information could get into the wrong hands, outside

of the US, especially when the employees working on that project have never intended to work on military projects and so are more willing to leak information.

In order to cooperate on a multilateral level, in May 2019 the OECD published its AI Principles and AI

Observatory, outlining principles and recommendations for governments in order to develop a level playing field. All 36 OECD countries together with a few others have signed the document. However, the US has signed the principles under President Trump, who

has continuously been expressing animosity towards international cooperation. China and Russia are only part of a consensus agreement stating they will support the efforts more broadly. 🚩

Lethal Autonomous Weapons Systems (see also: *Section 1: Defence and Cyber*)

An emerging and growing part of the foreign and defence policy discussion are lethal autonomous weapons systems (LAWS).

In 1988, a US guided missile cruiser shot down an Iranian passenger jet in the Persian Gulf, killing all 290 people on the plane. Even though the plane gave every indication to be a civilian airplane, the missile cruiser's Aegis system, programmed to target Soviet bombers, misidentified it. Nobody in the Aegis crew challenged the decision and so passively authorised the firing of the missile. A more recent example of LAWS is the Israeli drone 'Harpy'. It can stay high up in the air, observing a large radius of ground. When it detects a radar signal from the enemy, it crashes itself into the radar's location, destroying itself and everything around it.

When it comes to the fundamental question of life or death, it is questionable whether we really want to give a machine the full authority and control of that decision. In war, actions are time-sensitive and some might argue that these machines can take into account more information at a faster pace than any human could ever do. But the

decision to take away people's lives goes beyond purely rational calculations. War has become more complex and it has become more difficult to differentiate between civilians, enemies and allies. AI operated weapons are based on data, but what if that data is not sufficient? What if things have changed just the other day or hour and the programme is operating under false premises? Another argument against LAWS is one of responsibility and accountability. Who is responsible if something goes wrong and innocent people die? The scientist who programmed it? The commander who decided to use the weapon? The responsible government department which bought it? If responsibility defuses and the risk of being held accountable decreases, this can lead to decisions to kill people being made with less questioning. Lastly, because LAWS can be deployed with less risk to military personnel, the proliferation of such weapons might lower the bar for conflicts.

Following the moral arguments to ban LAWS, activists, over 110 non-governmental groups, the European Parliament, 26 Nobel prize win-

ners, more than 4,500 AI scientists and 30 different countries have joined a global campaign addressing the UN to ban LAWS. However, governments who drive the development of LAWS and profit from it have so far voted against the ban, which needs an unanimous vote in order to pass at the UN.

In the end, machines differ from humans as they don't have a moral compass. Morality itself is so complex and diverse that no machine will be able to be programmed with a moral compass. Guilt, shame, empathy, a feeling of responsibility and accountability are attributes that arise in a person when they do, see, or decide certain things. It will be very unlikely that machines, algorithms, or complex softwares replace these powerful human emotions. That is why it is so important not to give machines the power to make final decisions, or shirk from making decisions and taking actions just because the machine has chosen its course. An open, public discourse about morality in all aspects of life should take place to understand the distinction and uniqueness of humans and their difference to machines. 🚩

Bibliography

- "AI Policy - Singapore." Future of Life Institute. Accessed June 2, 2020. <https://futureoflife.org/ai-policy-singapore/>.
- "Artificial Intelligence." Infocomm Media Development Authority, January 22, 2020. <https://www.imda.gov.sg/AI>.
- Arnett, Eric H. "Welcome to Hyperwar." *The Bulletin* 48, no. 7 (September 1992): 14–22.
- Buchholz, Katharina. "China's Mobile Payment Adoption Beats All Others." *Statista*, May 7, 2019. <https://www.statista.com/chart/17909/pos-mobile-payment-user-penetration-rates/>.
- Can Kasapoğlu and Barış Kirdemir. "Artificial Intelligence and the Future of Conflict." Essay. In *New Perspectives on Shared Security: NATO's next 70 Years*. Accessed June 3, 2020. <https://carnegieeurope.eu/2019/11/28/artificial-intelligence-and-future-of-conflict-pub-80421>.
- "Chart: China's Mobile Payment Adoption Beats All Others ..." *Statista*. Accessed June 3, 2020. <https://www.statista.com/chart/17909/pos-mobile-payment-user-penetration-rates/>.
- "China Turns to Health-Rating Apps to Control Movements ..." *Statista*. Accessed June 3, 2020. <https://www.wsj.com/articles/china-turns-to-health-rating-apps-to-control-movements-during-coronavirus-outbreak-11582046508>.
- Clement, J. "Number of Internet Users in Selected Countries 2019." *Statista*, January 7, 2020. <https://www.statista.com/statistics/262966/number-of-internet-users-in-selected-countries/>.
- Conger, Kate, and Noam Schreiber. "California Bill Makes App-Based Companies Treat Workers as Employees." *The New York Times*, September 11, 2019. <https://www.nytimes.com/2019/09/11/technology/california-gig-economy-bill.html>.
- Cussins Newman, Jessica. "A Global Reference Point for AI Governance." Essay. In *AI GOVERNANCE IN 2019 A YEAR IN REVIEW*. Accessed June 3, 2020. <https://www.aigovernancereview.com>.
- Department of Defense. *Defense Advanced Research Projects Agency - Budget, Defense Advanced Research Projects Agency - Budget §*. Accessed June 7, 2020. <https://www.darpa.mil/about-us/budget>.
- "Digital Payments - China: Statista Market Forecast." *Statista*. Accessed June 7, 2020. <https://www.statista.com/outlook/296/117/digital-payments/china>.
- Elstrom, Peter. "China's Venture Capital Boom Shows Signs of Turning Into a Bust." *Bloomberg*, July 9, 2019. <https://www.bloomberg.com/news/articles/2019-07-09/china-s-venture-capital-boom-shows-signs-of-turning-into-a-bust>.
- Fang, Lee. "Google Hedges on Promise to End Controversial Involvement in Military Drone Control." *The Intercept*, March 1, 2019. <https://theintercept.com/2019/03/01/google-project-maven-contract/>.
- Franke, Ulrike. "Harnessing Artificial Intelligence." *European Council on Foreign Relations*, June 25, 2019. https://www.ecfr.eu/publications/summary/harnessing_artificial_intelligence.
- Galston, William A. "Why the Government Must Help Shape the Future of AI." *Brookings*. *Brookings*, October 25, 2019. <https://www.brookings.edu/research/why-the-government-must-help-shape-the-future-of-ai/>.
- Gansky, Ben, Michael Martin, and Ganesh Sitaraman. "Artificial Intelligence Is Too Important to Leave to Google and Facebook Alone." *The New York Times*, November 10, 2019. <https://www.nytimes.com/2019/11/10/opinion/artificial-intelligence-facebook-google.html>.
- Harwell, Drew. "FBI, ICE Find State Driver's License Photos Are a Gold Mine for Facial-Recognition Searches." *The Washington Post*, July 7, 2019. <https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/>.
- Kessel, Jonah M. "Killer Robots Aren't Regulated. Yet." *The New York Times*, December 13, 2019. <https://www.nytimes.com/2019/12/13/technology/autonomous-weapons-video.html>.
- Lin, Chia Jie. "Singapore Sets up AI Ethics Council." *GovInsider*, June 7, 2018. <https://govinsider.asia/innovation/singapore-sets-ai-ethics-council/>.
- Lin, Liza. "China Turns to Health-Rating Apps to Control Movements During Coronavirus Outbreak." *The Wall Street Journal*, February 18, 2020. <https://www.wsj.com/articles/china-turns-to-health-rating-apps-to-control-movements-during-coronavirus-outbreak-11582046508>.
- Lucas, Louise. "China Government Assigns Officials to Companies Including Alibaba." *Financial Times*, September 23, 2019. <https://www.ft.com/content/055a1864-ddd3-11e9-b112-9624ec9edc59>.
- McGee, Patrick. "How the Commercial Drone Market Became Big Business." *Financial Times*, November 27, 2019. <https://www.ft.com/content/cbd0d81a-0d40-11ea-bb52-34c8d9dc6d84>.
- Metz, Cade. "White House Earmarks New Money for A.I. and Quantum Computing." *The New York Times*, February 10, 2020. <https://www.nytimes.com/2020/02/10/technology/white-house-earmarks-new-money-for-ai-and-quantum-computing.html>.
- Mozur, Paul. "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority." *MinorityPaul*. *The New York Times*, April 14, 2019. <https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html>.
- Mozur, Paul, and Jane Perlez. "China Bets on Sensitive U.S. Start-Ups, Worrying the Pentagon." *The New York Times*, March 22, 2017. <https://www.nytimes.com/2017/03/22/technology/china-defense-start-ups.html>.
- Nuttal, Chris. "London Sets Standard for Surveillance Societies." *Financial Times*, August 1, 2019. <https://www.ft.com/content/70b35f8a-b47f-11e9-bec9-fd-cab53d6959>.
- Ray, Shaan. "History of AI." *Towards Data Science* (blog). *Medium*, August 11, 2018. <https://towardsdatascience.com/history-of-ai-484a86fc16ef>.
- Rosenbach, Eric, and Katherine Mansted. "How to Win the Battle Over Data." *Foreign Affairs*, September 17, 2019. <https://www.foreignaffairs.com/articles/2019-09-17/how-win-battle-over-data>.
- Sanger, David E., and Nicole Perloth. "Chinese Hackers Target Email Accounts of Biden Campaign Staff, Google Says." *The New York Times*, June 4, 2020. <https://www.nytimes.com/2020/06/04/us/politics/china-joe-biden-hackers.html>.
- "Some Aspects of UK Surveillance Regimes Violate Convention." *European Court of Human Rights*, September 13, 2018. *European Court of Human Rights*. [https://hudoc.echr.coe.int/eng-press#{"sort":\["kupdate%20Descending"\],"itemid":\["003-6187848-8026299"\]}](https://hudoc.echr.coe.int/eng-press#{).
- "The Campaign to Stop Killer Robots." *The Campaign to Stop Killer Robots*, 2020. <https://www.stopkillerrobots.org/>.
- "The Global AI Index." *Tortoise*. Accessed June 1, 2020. <https://www.tortoisemedia.com/intelligence/ai/>.
- "The War Against Immigrants - Trump's Tech Tools Powered by Palantir." *Rep. The War Against Immigrants - Trump's Tech Tools Powered by Palantir*. *Palantir and Mijente*, August 2019. https://mijente.net/wp-content/uploads/2019/08/Mijente-The-War-Against-Immigrants_-_Trump's-Tech-Tools-Powered-by-Palantir_.pdf.
- Wheeler, Tom. "History's Message about Regulating AI." *Rep. History's Message about Regulating AI*. *Brookings*, October 31, 2019. <https://www.brookings.edu/research/historys-message-about-regulating-ai/>.
- Wu, Tim. "America's Risky Approach to Artificial Intelligence." *New York Times*, October 7, 2019. <https://www.nytimes.com/2019/10/07/opinion/ai-research-funding.html>.
- Yang, Yuan. "The Chinese Internet Boom in Charts." *Financial Times*, August 21, 2018. <https://www.ft.com/content/ef80e27c-a500-11e8-8ecf-a7ae1beff35b>.
- Zwetsloot, Remco, Roxanne Heston, and Zachary Arnold. "Strengthening the U.S. AI Workforce." *Strengthening the U.S. AI Workforce*. *Center for Security and Emerging Technology*, September 2019. https://cset.georgetown.edu/wp-content/uploads/CSET_U.S._AI_Workforce.pdf.



Hannah Fuchs holds an MSc in EU Politics from the London School of Economics. She has written for several leading publications, including Policy Network, Euractiv, and Cicero. Currently, she is working as the communications officer at the tech and children's charity Lifelites.



SECTION 3

**CONCLUSIONS:
POLICY PROPOSALS**

INTRODUCTION

This Industrial Revolution, Britain must redefine the fundamental role of the state and the social contract. The key difference between AI and other forms of productivity increase is that it is autonomous, and thus a new tripartite contract between the state, humans, and human-like machines must be drawn up. We must accept this and embrace the vast opportunities opened up to us, but also manage the risks posed by this potential.

The debate around the tripartite social contract is only just beginning. The editor's view is that human-like machines (specifically, those who have individual human-level impact) need their own legal regime, but one that connects inseparably to the existing human one, just as human-like machines live inseparably with humans.

The proposals in this document are designed to both solve immediate problems but are also consistent with the above interpretation of the tripartite social contract. As we do more and more work in this area, so will precedents be set and a clearer vision be formed.

This section is based on three key insights which must be digested by policymakers:

1. We already delegate agency to AI even if nobody knows how exactly these, or the end-to-end processes in which they are embedded in, work
 - See Embracing the Potential: Proposals for an AI Regulator
 - News articles written by AI are used by market-making algorithms to set prices of everything from commodi-

ties and energy to medical supplies

- The concept of agency is process-dependent. If a doctor has to review AI-given suggestions at such a speed that it is not possible to provide independent opinions then she has effectively delegated agency even though she is “pressing the button” on the final decision. The same reasoning applies to lethal autonomous weapons (LAWs)
 - If agency is a line at which policymakers wish to draw, regulation needs to have control over process, as well as auditing the actual AI
 - Agency is crucial to define for insurance purposes. Proposed solutions, such as designated managers, are discussed in the relevant section
2. We have no clarity on existing treatment of data in our society
 - See Data Saves Lives - Consider it Vital Infrastructure
 - One example is the proposed anonymisation solution for privacy. Data can be de-anonymised by essentially cross-merging “anonymous” datasets
 - There needs to be a debate around whether data anonymisation is the most effective solution to data privacy. With anonymised or hashed data, it is difficult to detect biases and difficult to validate true anonymity
 - An alternative to anonymisation is regulatory control.

Personal characteristics, for example, can be collected specifically to check for bias. To enforce fair use, this must sit within an overall ethical framework

- How much is data worth to the individual versus the collective? Does the state have the right to forcibly collect data if it improves outcomes for the collective?
3. Parts of the current market are inefficient by the standard of a free, fair, and competitive marketplace
 - See Creating a Fair & Competitive Marketplace
 - Specific measures are required, but not sufficient, to enforce a fair market
 - Transparency: data lineage, showing an “ingredients list”
 - Enforcement: auditor with the power to penalise misuse
 - Anti-monopolistic measures: breaking up network effects in data or structure. Enforcing portability of certain data, such as social graphs in social networks, greatly reduces the barrier to entry, but come with associated data privacy challenges





Overall, it must be acknowledged that a more supportive government approach - a government able to execute the above measures and give clear answers to ethical questions - will confer economic benefits and develop the market faster, contributing to our reputation as one of the world-leading markets in this area.

Britain is one of the few countries with the ability to become a world

leader:

- Comparative advantages: world-leading universities, talent, attractive places to live and work, with a long history and reputation of free trade and welcoming investment
- A high-quality, sufficiently centralised state that takes decisions speedily but also allow sub-bodies enough independence to operate with technical competence - con-

trast this to the US and the EU

- Our democratic accountability and focus on privacy and individual liberties will require a market different from East Asia
- Britain remains an important world market which can't be ignored by any global enterprise

EMBRACING THE POTENTIAL: PROPOSALS FOR AN AI REGULATOR

A regulator specifically for AI is an idea actively debated amongst the political and technological community^{1,2} and most discussions on the challenges of AI land on some sort of regulatory proposal. The main alternative to a dedicated regulator is broad regulation (upskilling) across existing bodies. What is missing, however, are the details of how any regulator may operate and for what reason they need certain powers. This section sets out to fill that gap, and more importantly, how to get there from today.

“...we have concluded that the CMA’s current tools... are not sufficient to protect competition... We therefore recommend a new pro-competition regulatory regime with strong and clear ex ante rules, which can address a wide range of concerns holistically, can be enforced rapidly by a dedicated regulatory body” - CMA Online platforms and digital advertising market study, July 2020³

Consider a new model of aeroplane which nobody knows how to operate. It might flap its wings, float like a balloon, or inflate itself. It works, but nobody knows how. Even the people who built it do not know. What we do know is that it’s much faster and safer than a regular plane. Would you fly?

The above anecdote illustrates the inefficiency of a market where the products cannot be tested by the consumers. Even if consumers cannot directly test the safety of planes, there is a robust international framework, run by experts and ultimately accountable to the people, which governs their safety and builds trust with the consumers. Currently, the usage of AI is almost invisible to the everyday consumer and there is no oversight to

monitor, say, if crucial anti-discrimination laws are being broken by decision-taking algorithms.

One regulator, set up correctly, would be better than broad regulation (upskilling all existing regulators) due to economies of scale and fairness - ensuring auditing, for example, is consistent with one ethical framework across all use cases in the country. We cannot risk certain industries being left behind due to regulatory capacity and AI regulatory talent - highly in demand on the best of days - being spread thin across the almost 100 individual regulators in the UK. Crucially for business, this regulator must be set up without overlap with existing regulators. There will be industry-specific questions (are maintaining bank capital buffers considered “critical” enough to warrant extra oversight in that area?) that the single regulator will have to work with, but the pool of experts who are forming a consistent ethical framework, the foundation of the tripartite social contract, is better concentrated in one body.

Current data and AI governance is already splintered across industries and we already live with inconsistencies across the country. Again, this is not necessarily the fault of any specific institution, but more a reflection of how difficult it is to build and retain expertise in the area. The financial services industry enjoys the benefit of a well-funded, globally respected and forwards-looking regulator in the form of the FCA, which together with the Bank of England published a report in Oct 2019 on the use of AI in the industry⁴ and is actively looking at industry use cases. The CQC, on the other hand, is one body near the other end of regulatory preparedness and is only just

now looking at data privacy issues. Local government procurement and data sharing regulations are also such that obtaining approval for many digital schemes are already prohibitive, which does not portend well for the future.

It is worth noting that an independent regulator is not the only way of protecting citizens’ rights and safeguarding a fair market. A new government ministry, with a dedicated minister more directly accountable to citizens, may be required to make inherently political decisions around issues such as data privacy.

A summary of the reasons why we need such a body:

- Auditing: making sure AI works fairly for all, and is compliant with existing laws (e.g. anti-discrimination). Auditing is necessary to protect and validate existing citizen rights, including but not limited to privacy, market access, and discrimination. As discussed above, end-to-end, contextual process monitoring must be a part of the auditing powers
- Creating a fair marketplace and promoting competition: the power to investigate and break up network effects, enforce transparency in data markets, and ensure consumers are being offered fair products and services
- This is discussed in further detail in the following sections
- Civil rights: an arbitrator to settle disputes, apply restitution, quantify damages (specialist support for the legal system). Help develop

- a new insurance and legal framework for AI
- Manage cases of citizen abuse e.g. algorithmic racism & sexism
- Encouraging investment: ensure that the UK market is a safe one to invest in, and clarify the boundaries of the UK “ethical market” to create an investable one in the first place
- ESG funds are driving investment flows - over \$1tn at the time of publication⁵ - and tech stocks are under scrutiny over their existing data privacy and ethical uses⁶, indicating that clear ethical AI companies would attract investment
- Advise on ongoing developments in the industry as well as encourage industry research in particular areas e.g. how to respond to adversarial attacks on AI
- Promoting the UK’s status as a world leader by being one of the first countries to take this step. We know it is possible for us to lead a market like this by example of the Bank of England and the FCA, which are globally respected institutions
- Promoting operational efficiency and security within the public sector: this is further discussed in the Data section

How should such an institution function to be most effective?

- Using a standards-based, rather than legalistic, model which relies on ethical and political principles to ensure accountability. This model remains flexible to quickly cover grey areas, which is important in an emerging field
- By being reactive, monitoring, and proactive - looking at the entire development lifecycle
- How precautionary to be is a political choice:
 - Features from other regulatory models can be borrowed - for example, the ICAO’s best practices in error tracking and learning from mistakes, or an FDA-style “stamp” of pre-market approval may be applicable to “ethical markets” such as health, if a more conservative approach is desirable
- Monitoring of the market via auditing the applications of AI in the context of their surrounding processes
 - The regulator will likely have to prioritise top-tier impact uses of AI, along with companies which have received the most complaints
- “Bug bounties” can be used as incentives as part of ongoing monitoring and enhanced auditing, especially for non-sensitive use cases⁷
- Proactive and forwards-looking with an aim to be world-leading and influential. This part is crucial to becoming a leading market force. Otherwise, a regulator can end up being purely reactive and seen to be holding back the market
 - Principles such as being iterative, outcome-based, decentralised, and inclusive. The innovation foundation Nesta’s work on “anticipatory regulation” is influential⁸
 - The FCA’s regulatory sandbox, a structure used to test out new innovations in a safe space, has been well-received by the industry. An equivalent sandbox for AI would create a space for innovative companies to experiment and develop
- The ability to levy penalties for non-compliance. Conversely, if standards can be created for areas such as data quality, these will act as soft positive incentives if these standards are adopted by the industry. They can be, for example, prerequisites for government procurement
- Tiered regulation by industry sector, with higher tiers subject to more scrutiny and a higher priority, and tiers dependent on impact on human lives. Health and financial services will likely be at the top, with other consumer applications (e.g. product recommendation systems) lower down. This clarity will help safeguard and ringfence the sensitive and ethical market for the private sector to develop solutions for, whilst preserving the exciting innovation in other areas of the economy
- For top-tier impact cases, additional measures such as the publishing of impact assessments of ethical considerations and potential societal consequences can
 - a) make easier auditing and
 - b) force designers to take them into account during the development process
- Assigning accountability for legal purposes: what happens if an AI causes harm?

- A solution is “accountable dedicated managers” for critical AI applications, who are legally liable for prosecution
- Official oversight over government use of data and intra-governmental data transfers. “Data corridors” between government entities need to be governed appropriately and set up where they don’t currently exist to increase operational efficiency of government
- Using the latest technology to reduce compliance costs. Look into solutions such as creating secure channels (APIs) to directly pull data from companies for reporting and auditing
- High diversity standards for its own management board - because important regulatory decisions which can impact the entire market should fairly represent the whole country’s population

What is the Government currently doing?

The Information Commissioner’s Office, as the UK’s independent information rights regulator, is perhaps the best placed body to be given a broader remit involving AI.

The ICO has been recently conducting work on putting together a draft AI auditing framework⁹. This, of course, is a good starting point. However, the Government needs to reach much further and broader if a truly effective regulator is to be formed to bring all the benefits outlined above to the market. Audit is just one of the many pieces needed to be put together to achieve a desirable outcome.

There have been a few other announcements, the main item of interest being a proposed “Big Tech Regulator”¹⁰ in 2020, following the Furman review¹¹ and Digital Competition Expert Panel¹² studies. However, detail is very much missing from the public sphere and there is no clarity on what the exact remits of these regulators will be. Some considerations are:

What is the remit of the regulator? Will it audit AI itself or only deal with competition?


If the remit is only big tech (the term used is “Strategic Market Status” companies), the institution will be reactive to one specific part of the ecosystem rather than a proactive and monitoring institution that we need for the whole economy. Additionally, there will be a vulnerability to regulatory capture if only big tech is engaged and not other parts of the economy

May do little for SMEs not already affected by big tech. Doesn’t provide an avenue to help boost productivity

Proposal seems to be more aligned to a “Digital CMA” specifically for big tech rather than a single body with remit for all AI related applications

More importantly, the National Data Strategy is delayed, having previously been promised by the Government for 2019. Without

this, we have no clarity on how important the UK’s data is for the Government, and therefore no direction on how to resolve existing data concerns such as privacy, anonymisation, and national security.

A recent Committee for Standards of Public Life (Feb 2020) report¹³ concludes that the current status quo is too immature for a regulator. This viewpoint, which is more from practical rather than ideological concerns, makes a point about a missing framework. However, the formation of a regulator can precede the formation of a framework (as the Government have announced with their “Big Tech” regulator) and can assist in establishing such a framework by precedent. A regulator can already start working on research, set up vital links within the Government, and start auditing data and AI usage within the public sector and companies. The establishment of a body is also a signal to the international community that the UK is serious about being a world leader. 

DATA SAVES LIVES - CONSIDER IT VITAL INFRASTRUCTURE

What is data?

How should we treat data as an economic resource? We already know that efficient and responsible use of data saves lives, allows businesses to thrive, and removes a lot of pain from people's lives. However, the way we look at data can greatly affect what we are comfortable doing with it.

Various thinkers compare data to commodities¹⁴ or capital¹⁵, but considering data as intellectual property gives one of the best frameworks for understanding what we need to do to use it responsibly.

Public data, referring to impersonal public goods such as a geographical map of the UK, climate and air pollution data, and atmospheric data, can indeed be largely treated as a public good or reusable commodity which in some cases is best invested into by the state due to the economies of scale. (The editor uses the term "public" in this phrase referring not to privacy but to the economic definition of a public good.)

As IP is considered intangible property as a result of creativity, so should personal data be similarly considered intangible property as a result of socially-defined privacy structures. As we have seen throughout this pamphlet, privacy rights are a globally varying construct. This is why the Government's National Data Strategy needs to urgently clarify the UK's stance, but this will likely not answer several crucial questions, one of the most critical being the right of the government to forcibly collect data from its citizens.

Is forced data collection the new cost of living we must accept, or is data privacy a fundamental human

right? The current landscape has many commentators arguing for the latter. However, the advantages of a theoretically perfectly secure database of individual data, accessible only to researchers and the benefit for the public good, are undeniable. An incredible amount of medical insight and scientific progress could be gleaned and used to find treatments and extend life expectancy for the collective. Parts of the economy requiring support could be identified and efficient, targeted support provided to certain regions or sectors. Tax avoidance could become an issue of the past. The strength of the common endeavour achieves infinitely more than can be achieved with individual data points, and ensures that there is more accurate representation of minorities compared to voluntary data collection. There is, essentially, no fight between "my data" and "your health", but my data helps improve your health, and your data helps improve my health. Of course, such an implementation is beholden to practical challenges, but the idea of utilising our collective data as a national resource for the benefit of all should not be shied away from.

Comparing data to IP yields the following insights:

- Data takes effort to gather and curate, and is done so for specific purposes. It is not a true free public good (e.g. the ocean) and is highly context-specific, unlike commodities
- The economics on the "return" on data, like there is on the return on capital, is not well understood. There are specific network effects but these rely on other technologies in addition to the data to generate a "return" (e.g. search engines). What is clear, however, is that data doesn't conform to any well-understood economic model and attention needs to be paid to emerging research in the area to fully understand it. Existing models must be treated with caution
- Data is not transferable and taxable like capital. Due to the value of data being context-specific and future value flows unpredictable (unlike universal capital), it is not possible to assign a present value and therefore tax it. Additionally, there is no equivalent to "free flow of data" because of its replicability and privacy constraints
- There needs to be sufficient incentives for data gatherers, especially in the case of wishing to incentivise creation of public data
- In the case of the state gathering personal data, there should be incentives for data owners to accede to giving up their data. What this whole bargain is needs to be defined in the tripartite social contract
- A 2019 paper in the National Bureau of Economic Research¹⁶ models data sharing in the market and finds that allocating property rights to consumers for their data results in more optimal market allocation (see more detail in Creating a Fair & Competitive

Marketplace)

Data is a “raw input” for industry - especially new B2B companies who use it to boost efficiencies or enhance existing services - entire companies (such as Reuters) exist purely to process and sell data. However, these depend heavily on context, again, like IP. Data depends heavily on use cases - it is never usable out of the box like capital or commodities - and the concept of data belonging to a data owner conforms with our notions of individual rights, rather than simply being a resource which is “gathered”.

Privacy and ownership of data are also concepts which need to be defined by the state via the citizenry. It may seem like privacy is a simple concept - what’s mine is mine - but for replicable, social data, there are plenty of grey areas in which the rules are unclear. To whom do the timestamps of messages sent over a messaging platform belong to? Who is allowed to know the people I know - the “social graph” in social networking platforms? What rights do I have to the pictures of me, even if I do not know they exist?

Data anonymity is also a concept which needs to be challenged as we come to understand the deductive capabilities of social data. A study in 2019 by leading privacy professionals at Imperial College London and UCLouvain showed that 99.98% of individuals could be correctly re-identified in a dataset compliant with modern GDPR and CCPR standards¹⁷.

The challenge with “anonymous” data as a term in discourse is that there are neither universal standards nor a complete mathematical understanding of the attributes that make data anonymous, further compounded by the fact that definitions of anonymity must rely heavily on social definitions. Contrast this to, say, encryption, a much better understood mathematical problem which enables accurate estimates of security and thus other technology which relies on it. Like GANs (see *Section 1: Communications*), we must accept instead that a technological arms race is happening in this space, and look for policy solutions which aim for greater control of data, assuming in the worst case that data

cannot be made truly anonymous. The technology is not yet there to assure that data privacy is a fundamentally achievable goal, so we must plan around it.

The previous paragraphs are not just important in clarifying individual and business rights, but must be answered to solve the data interoperability puzzle. Governments, in this debate, must represent the people’s interests along with private companies, who are currently leading the discussion in this area¹⁸. Business is finding the theoretical concept of interoperability, supported by GDPR and other regulations, difficult to realise in practice due to barriers and multiple standards in data organisation. Here, governments can help by helping design standards for certain types of data, such as SWIFT does for international payments, resulting in much quicker data transfer and real portability. Portability is, many would argue, a crucial piece of the puzzle to breaking up network effects (see the section below, *Creating a Fair & Competitive Marketplace*). 🚩

Acknowledging investment needs

In the March 2020 budget, £2.5bn was announced for fixing potholes¹⁹, but the only references to data were £16.4m over three years for data sharing within government, and £5m for economic data.

There was no recognition for the need for data centres, technical infrastructure, open source data, and ensuring a competitive digital marketplace - in other words, the new vital common goods which our knowledge-based economy and regular businesses need.

Data must be classified and treated as vital infrastructure. This semantic step is critical to prioritising the infrastructure that underpins our modern, knowledge-based econ-

omy. Our response in crises such as COVID-19, or even war, could be greatly improved with truly modern digital infrastructure. Regular businesses benefit strongly from safe, well-structured internal data infrastructure which unlocks opportunities and trade. Conversely, loss of human life can be attributed to a direct result of negligence in this area.

Practical steps can be taken on the back of this. The first is to add data infrastructure investments into the definition of R&D expenditure, so that public and private (crowd-in) investment can benefit from government support for improving their back-end systems.

Various announcements on pub-


lic sector investment into “infrastructure” have not come close to touching digital infrastructure in a meaningful way. Underinvestment in an economy which is 80% services based²⁰ means that a very significant portion will be related to digital infrastructure.

Forward-thinking financiers are already identifying good opportunities for their capital which also advance this national industrial strategy, a triple win for business, financial services, and the country. However, the financing gap in the UK²¹, reflecting a structural investment gap, shows that we are still taking too long to truly adapt and help solve the productivity puzzle, holding our businesses back. A

push for experienced technology bankers in an institution such as the British Business Bank could stimulate competition whilst preserving a fair market (the BBB does not break EU state-aid rules). Of

course, a full state NIB (with a legal mandate and full government guarantee) is always a more powerful option, but is less able to be directed by government policy.


Innovative financing may be another

solution which can emerge by itself with enough competition in the marketplace - for example, "data bonds" as a complementary ESG or strategic financial asset. 

Acknowledging investment needs

Although the ICO is doing good work already, more needs to be done in preparation for the current age of widespread AI.

Data is the foundation upon which AI is built upon and policy involving data must always be considered alongside AI policy.

- The AI regulator or the ICO needs a bigger mandate to monitor data after we designate it critical infrastructure
- Hold the public sector to standards involving data preparedness by monitoring key metrics and stress testing
- Oversight of data corridors which run between different parts of the public sector to improve operational efficiency. A review of all intra-governmental data transfer points will yield many areas in which operational efficiencies can be found, processes automated, and security enhanced
- Ability to audit the private sector must line up to the National Data Strategy. The Strategy will define how important general private sector preparedness must be, and the auditing and penalty powers must be sufficient to enforce this
- As explored above, if public standards can be formed for data quality (likely based on architecture auditing), auditors can assign ratings to businesses
 - Standards can act as soft incentives - for example, limiting government contracts to certain standards
 - If businesses require financing to improve poor data quality they can access "data bonds" which may be purpose-limited loans. The Government can choose to back these loans to further support the initiative
 - Financing made available to improve data security, reliability, lineage etc. available aligning to a National Data Strategy, which again needs to clarify the value of the nation's data from various angles, including a national security perspective
 - Examine the options to improve financing:
 - Direct government financing: improving financing using existing channels. Just £957m of new commitments were made by the British Business Bank in 2019²². Contrast this to the €36bn invested in German SME and Private Clients by KfW, the German state development bank²³
 - The formation of a full national investment bank, which requires a legal mandate and a state guarantee. This adds contingent liabilities to the state but not public debt
- Add data infrastructure investments into the definition of R&D expenditure, so that investment can benefit from government support in the form of tax credits
- Public data - continue targeting the #1 spot on the Open Data barometer²⁴, leading the world in procuring genuine "public good" datasets for individuals and businesses, and improving the quality of public data (open licenses, identifiers etc.)
- Private data - examine the potential benefits of a national database of citizen data, paying particular attention to the security and governance model that would govern it
 - Learning from the shortcomings of the care.data programme²⁵ would be a beneficial first step
- Programmes to help companies, particularly SMEs, with data management and data architecture and design
 - Can be like HMRC's existing business support programmes
 - Measure the efficacy: combine hard and soft targets which align with the National Data Strategy to ensure targets are met. For example, engagement by region / number of SMEs reached / qualitative surveys indicating helpfulness 

CREATING A FAIR AND COMPETITIVE MARKETPLACE



In the previous sections, we discussed the unique nature of data and the subsequently unusual economics that arises. It goes without saying that the current market is clearly imperfect, and measures can be taken to improve competitiveness.

The current system in the UK is particularly reactive (rather than proactive) to potential rule-breaking. Only big breaches, such as the Facebook-Cambridge Analytica data leak in 2018, are investigated

by the authorities, and then only after they arise. There remains little transparency in many other parts of the market which are predominantly data and AI-focussed. For example, 15% of online ad spending is untraceable, according to a recent two year long PwC study²⁶. The exact size of the market is difficult to measure, and estimates range from £2bn in the UK in this study to £14bn as estimated by the CMA²⁷.

The value of consumer data is

huge. A 2020 US Federal Reserve paper places the growth of consumer surplus, or value, of digital goods and services at \$30,000 per capita from 2004 to 2017²⁸. Taking this number at face value, it follows that the value of the consumer data which is provided to these platforms (which include free services such as Facebook and Google) ought to be a considerable percentage of that surplus. 🚩

Why is the market imperfect?

Economic research into data is in its preliminary stages. This section summarises our understanding so far, which goes some way to explaining the current market.

From a basic transparency standpoint, illusion of control over data acts as a significant barrier to achieving market equilibria²⁹. Forcing companies to disclose data usage in privacy policies, for example, has done very little to give users proper control over their data usage. An auditing body would do a much better job at spotting abuses - which is effectively collective action.

The network effects in data are vital to understanding its impact. A “normal” trade would have the buyer and seller each assessing the val-

ue of the product to them and thus arrive at a fair price. Data, however, greatly increases in value the more an actor has of it (both because of the inferences that can be gained from it and because a “complete” dataset is worth much more than an incomplete one), which means that its value is more than the sum of its individual parts.

Initial research into both the data purchasing market and the data market show significant challenges with the status quo (essentially a free market). A free data purchasing market is shown to have multiple, sometimes infinite, Nash equilibria^{30 31}, indicating a fundamental unpredictability in how the market operates, and recent models of a free data market show that it does not lead to an efficient use of information, nor privacy protection for

the consumers³². Data unions are a proposed hypothetical solution, involving collective bargaining by consumers which would more accurately capture the collective value of the data, but there exist significant implementation challenges³³.

Data leaks and network effects decrease the value of data to individuals due to privacy externalities³⁴. In simpler language, the value of data to the user (a secret, for example) decreases after the first time the user shares it. However, the value of that data usually remains the same to the company, creating a situation where asymmetry of information benefits the producer, who can acquire the data for less than they would otherwise have been able to. 🚩

What can we do to improve it?

The below policy proposals are aimed at countering the negative effects of a free data market as explored above:

- Data lineage: “ingredients list” showing the lineage of data processing, increasing the transparency of the marketplace
 - Implementation is flexible: there is a choice to force it to be shown for specific sectors, adverts, and showing specific corporations, parent companies, and could be linked to the existing Companies House database
 - Could show metadata, companies who have touched the data, or something else
 - Technical study required to gauge feasibility
 - Measures to make interoperability more functional
 - Launch a consultation with both big and small companies to see what improvements to interoperability³⁵
- are required to make it easier to use data
- Consider new funding models for data procurement to ensure that incentives align for producers
 - The patent model may be a potential solution to data markets. This can apply both ways - private datasets can be protected for time-limited periods, increasing incentives to procure them, or if states wish to protect, track, and/or gain value from its datasets, they can license the usage
 - A subscription service may be another way governments can incentivise production, in cases which datasets need to be continuously procured. This can be similar to the NHS trial for procuring drugs to defeat antimicrobial resistance³⁶
 - Network effects (data gathering from product use is a positive externality which produces a better product which gathers even more data) can be broken up by targeting specific parts of the externality-generating product
- Anything that relies on social networking, for example, can be made more competitive by making the social graph interoperable, leaving the UX and content algorithms open to competition
 - Similarly for search engines, web indexes and user feedback metadata are the main barriers to entry
 - Specific suggestions for big tech have been suggested by the CMA in its interim report³⁷
 - Consider collective bargaining mechanisms for consumers of services where data is given in exchange for products
 - Explore the possibility of “data unions” as mentioned above

OUTCOMES OF A RESPONSIBLE REGULATORY MODEL

What would a medium-term future look like for the UK economy with the above series of policy proposals implemented?

The UK could take one of two broad paths: continue along the lines of its global free-market model, or allow the state to pursue a greater role in economic development.

The current geopolitical situation makes the former seem riskier than it was just five or ten years ago. The success upon which this type of economy is built upon depends on a rules-based international order being upheld, and

recent events have attacked that confidence. That being said, there are still benefits the UK could gain from a progressive regulatory regime, including a fairer market with greater competition and greater welfare for consumers, the ability for consumers to seek redress, and a more efficient government machine.

A more active state role unlocks a host of new possibilities for the UK. If the state were to assert full oversight over the UK's data, it would see national data, both public and private, conform to a National Data Strategy. The entire economy would be revamped to

fully support a data-driven economy, with much bigger investments in infrastructure, R&D, and education (both in the formal sense, likely in the form of state-sponsored PhDs, and business outreach). On a global level, the UK could be aggressive in attracting talent from all over the world with visa exemptions and tax breaks. Providing a national source of financing as an alternative to global takeovers of British firms would also be a priority as data and AI companies become more important when seen through a national security lens. 🇬🇧

CONCLUSION: FACING THE FUTURE

We are all living at an inflection point in human history.

All human creations are a product of intelligence. Never before has a technology appeared to challenge that human monopoly; to make intelligent decisions for other human beings.

What we must decide now is how, not if, to live alongside human-like machines. Refusing to do so would be turning our back upon progress, and all the potential to make human life better and easier.

The fear of this technology is somewhat justified. But the fear should be of human mismanagement, not the technology itself. From poor governance causing harm to malicious actors exploiting AI for their own gain, it is human social struc-

tures which must adapt to accommodate this new category of being



- the human-like machine.

There are no politicians who advocate for a complete lack of rules governing human beings. So why would we let human-like machines exist in the same anarchy? Harms resulting from lack of oversight are our own failures, and can dent public confidence and progress in the

beneficial uses of the technology.

An “ethical framework” or “regulatory regime” is not ambitious enough. We must reconsider the entire social contract and rewrite it to include human-like machines. The policy ideas in this pamphlet will go some way to beginning this monumental task by sparking debate around ideas not yet considered.

We look forward to continuing the discussion of the tripartite social contract in further work. 🇬🇧


Marcus Storm
 Editor

Marcus Storm heads AI products for a business line at a global investment bank. His aim is to make lives better through politics and technology. He speaks eight languages and engages in politics internationally. Find out more about his work on his website, www.marcus-storm.com

GLOSSARY

As with any field, technical jargon can present a barrier to understanding the latest developments. In the case of AI this is made worse by the facts that: i) the field is moving quickly; ii) the concept of AI itself is not universally agreed upon; and iii) many companies and individuals often misuse this jargon when marketing themselves amid the current wave of AI ‘hype’.

Below we have defined some of the key terms and acronyms referred to throughout the pamphlet. In general, when referring to AI, we refer to Narrow AI and its subfields which are widely in use today.

- **Artificial General Intelligence (AGI)**, in contrast to the **Narrow AI** we have today, refers to a theoretical machine intelligence that is as capable as a human (or more so) in performing any intellectual task.
- **Artificial Intelligence (AI)** is, roughly, the ability of a machine to perform intellectual tasks that one would typically assume to require human-like cognitive abilities such as perception or logical reasoning.
- **Autonomous Vehicles (AVs)**, for example self-driving cars, are vehicles that are equipped with various technologies (many using AI) that allow them to drive with little to no human input.
- **Bias** (in the context of data and AI) refers to systematic errors or misrepresentations in datasets or algorithms (possibly created using **Machine Learning**) that may lead to unfair disparities in how different groups are treated.
- **Big Data** is a name for incredibly large, complex datasets characterised by their volume, variety, and velocity (this third adjective describing their growth in size).
- **Blockchains** are not a form of AI, but simply a cryptographically secured list of data records often used to encode the distributed ledgers of **Cryptocurrencies** such as Bitcoin.
- **Computer Vision** is a subfield of AI that attempts to automate tasks relating to the processing and interpretation of images and videos, such as object recognition.
- **Cryptocurrencies** are digital currencies whereby individual ownership is recorded cryptographically on distributed ledgers, such as **Blockchains**.
- **Cybernetics** is a (somewhat outdated) name for the transdisciplinary study of control, regulation, and dynamics in information processing systems, both natural and artificial.
- **Cybersecurity** is a field that seeks to protect hardware, software, and data from theft, damage, or other disruption due to malevolent actors and algorithms.
- **Data Mining** refers to a set of methods for extracting patterns and information from datasets using tools from computer science, statistics, and increasingly AI.
- **Data Science** is an independent field from AI that draws on interdisciplinary methods, such as **Data Mining** and **Machine Learning**, in order to extract knowledge and understanding from data, often

By Lewis Hammond

Big Data.

- **Deep Learning** is a subfield of **Machine Learning** that relies on deep Neural Networks to approximate complex functions which can then be used for a variety of purposes.
- **Evolutionary Computation** is a subfield of AI studying methods of computation inspired by evolution, such as having a population of candidate solutions to a problem which are then updated over time and evaluated by a ‘fitness function’.
- **Expert Systems** are AI systems comprising a knowledge base containing facts about the world and an inference engine which uses this knowledge to answer queries and deduce new facts.
- **Generative Adversarial Networks (GANs)** are formed by a pair of **Neural Networks**, one of which learns to generate realistic data (such as images or videos) by attempting to ‘fool’ the other into thinking the new data is part of the original data.
- **Good Old-Fashioned AI (GOFAI)** is a (somewhat pejorative) term for older approaches to designing AI systems based upon logic and symbolic reasoning, which was the dominant paradigm for the first 30 years or so of AI research.
- **Intelligent Agents** are autonomous entities that pursue goals in some environment by perceiving information and taking actions accordingly, based upon some reasoning process.
- **Knowledge Representation and Reasoning (KRR)** is a

subfield of AI focusing on the design of computational representations of information, along with processes that may be used to generate or deduce novel information using these representations.

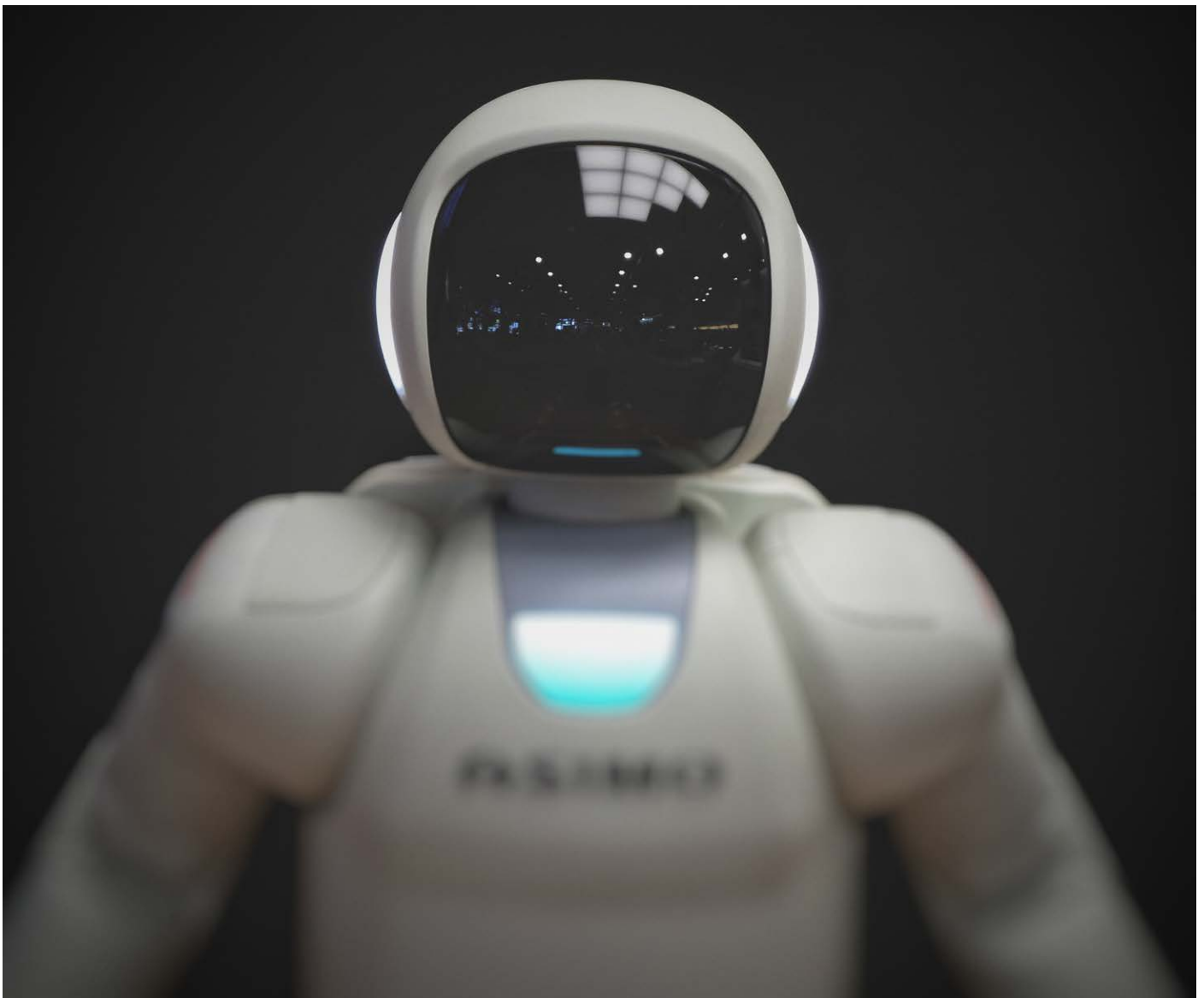
- **Machine Learning (ML)** is a subfield of **AI** in which computational models are created using algorithms which learn from experience (known as 'training data') rather than being explicitly programmed.
- **Narrow AI** (sometimes also called **Tool AI**) refers to **AI** systems that are created in order to perform specific (i.e. narrow) tasks as opposed to having the general reasoning capabilities that humans possess.

sess.

- **Natural Language Processing (NLP)** is a subfield of **AI** which seeks to automate tasks to do with natural language (those that humans use, as opposed to the formal languages of logic and computer science) such as speech recognition and translation.
- **Neural Networks (NNs)** are biologically inspired **Machine Learning** models containing multiple layers of small computational units ('neurons') connected to one another using simple functions, which together approximate more complex functions.
- **Planning** is a subfield of **AI**

that develops algorithms for creating strategies and plans that solve complex tasks, which are often then executed by **Intelligent Agents**.

- **Reinforcement Learning (RL)** is a subfield of **Machine Learning** in which **Intelligent Agents** learn to perform tasks in unknown environments by seeking to maximise 'rewards' that are given to them when performing well.
- **Robotics** is an independent field from **AI** at the intersection of engineering and computer science which develops and studies robots – machines that act autonomously to perform complex physical tasks. 🚩



DISCUSSION GUIDE

How to use this discussion guide

The guide can be used in various ways by Fabian local societies, local political party meetings and trade union branches, student societies, NGOs and other groups.

You might hold a discussion among local members or invite a guest speaker – for example, an MP, academic, local practitioner, or one of the writers to lead a group discussion.

Some different key themes are suggested. You might choose to spend 15–20 minutes on each area, or decide to focus the whole discussion on one of the issues for a more detailed discussion.

A discussion could address some or all of the following questions:

1. What were your views on AI before and after reading this pamphlet? How have they evolved and why?
2. This pamphlet focusses on the key sectors of the economy which will be transformed by AI in Section 1. What are your thoughts on these areas – for example, in education, policing, and health? Imagine if current jobs, like police, teachers, and nurses, were replaced by AI. How do you feel about that prospect and why?
3. Section 2 focusses on international competition. How do you think this will play out in the future? Will friendly economic competition in this space turn into something else, or be affected by national security factors?
4. The conclusions in Section 3 lay out two contrasting visions of responsible regulation – a more non-interventionist outcome and a more active state. Which one would you like to see and why?
5. Marcus Storm introduces the concept of a tripartite social contract – a settlement between citizens, the state, and human-like machines. What are your thoughts on this? How would you like to live alongside human-like machines?

Please let us know what you think

Whatever view you take of the issues, we would very much like to hear about your discussion. Please send us a summary of your debate (perhaps 300 words) to publications@marcus-storm.com

In **Modern Britain: Global Leader in Ethical AI**, we cover the use of AI in a huge range of applications across the entire economy and how the world is reacting to it. What links all these disparate threads together is the autonomy of the technology. The ability to make its own decisions. In short, the intelligence.

It is in all our interests that Britain takes a global lead in AI. A new National Industrial Strategy must be formed on the back of a tripartite social contract – a new settlement between the state, citizens, and human-like machines. Harms caused by AI are ultimately failures in human governance which limit the vast potential to improve lives all around the world.

Modern Britain: Global Leader in Ethical AI brings an astounding variety of new ideas to the fast-paced and deep debate around AI and ethics. Now is the right time to bring the debate to living rooms across the country to ensure that we all understand the stakes and have a voice in the future we will have to share – with the machines.

Edited by Marcus Storm, with contributions from Darren Jones MP, Ivana Bartoletti, Tom Grand, Kamal Puwar, Anita Chandran, Mohamed Hammeda, Cecilia Eve, Luke Richards, Kyran Schmidt, Hannah Fuchs, Lewis Hammond, Antoinette Hage, and Tom Ascott

