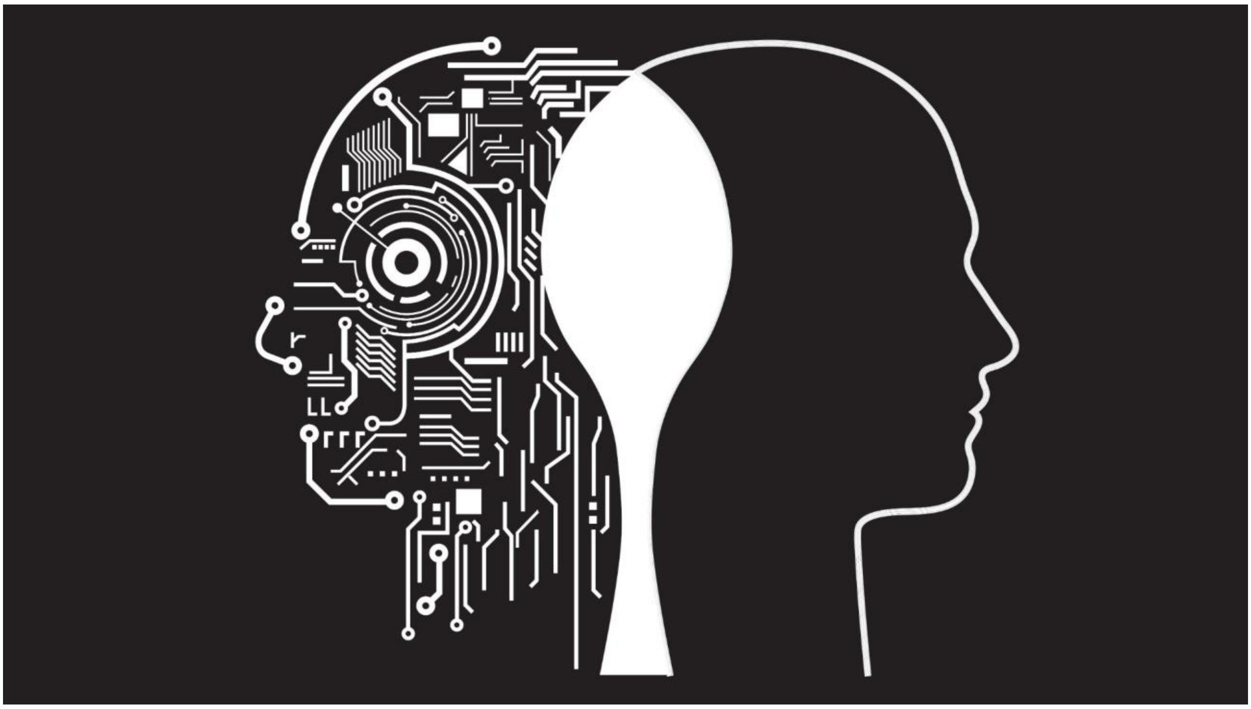
THE WILBERFORCE SOCIETY



Navigating AI through the 21st Century

Editor: Daniel Hurt

Writers: Beth Barnes, Riccardo Conci, Sobia Hamid, Daniel Hurt, Ed Leon Klinger, Gregory Lewis, Cameron Wallace

Advisors: Shahar Avin, Clemens Öllinger-Guptara, Chia Jeng Yang

EXECUTIVE SUMMARY

This paper offers some suggestions on how governments and businesses can manage the risks, whilst maximising the benefits, posed by Artificial Intelligence (AI) over the course of this century. It begins by outlining the history and present state of the art, summarising predictions made by experts in the field on how it might progress in the coming decades. Following this it provides an overview of the dangers that advanced AI applications may present across a range of industries.

It then conducts a more detailed analysis of these areas, each with its own policy recommendations. The topics we look at in greater detail are the following:

Financial Services:

Retail banking and high finance are two services where AI is already prolific. Who are the relevant individual and institutional stakeholders? How can they be protected from market and technical failures originating from AI applications?

Medicine:

AI could be used as a complement to human health workers, particularly in the primary care sector. How can we fully realise the potential of AI-based clinical decision making for aiding in the diagnosis and management of diseases? What should we do to deal with ethico-legal issues of smart wearables?

Autonomous Vehicles:

Autonomous vehicle technology presents immense benefits to road safety, reduces emissions and compliments the electrification of automotive transport, and offers a life-changing option for elderly or disabled people who are currently unable to drive. As with many disruptive technologies, it presents risks as well, and we ought to take steps to mitigate such risks. How will industries be affected? What safety risks are associated with driverless cars, and how can we combat them?

Greater Than Human Intelligence:

> Machines that match or exceed human intelligence could be hugely beneficial or catastrophically destructive. What needs to be done to reduce these risks whilst realising its outstanding promises?

In its conclusion the paper provides a summary of the policy recommendations made across these areas. It continues with a discussion of the policy themes that have been developed, followed by some final remarks. Throughout we avoid discussion of the technical details of AI development, but focus on how states and private institutions can increase the likelihood that the social, economic and political impact of advanced AI will be positive. To this end it makes recommendations in the following themes:

- Develop flexible anticipatory frameworks across all industries discussed that will facilitate adaptation to rapid or unanticipated progress in AI. These must be drawn up in collaboration with experts who are most knowledgeable about the future of the field, and designed to ensure developments are socially beneficent.

- Improve dialogue between researchers, industry, and government in order to foster a culture of openness and safety. Not only will this help to avoid undesirable outcomes of AI in certain areas where it poses dangers to human life, it will help policymakers to lay the groundwork for fully realising its positive effects for their constituents.

- Encourage international collaboration on AI, particularly in situations where it might be used for military purposes. As we approach human-level machine intelligence in future decades this may be essential to avoid catastrophic outcomes. It is also important in regards to sharing the benefits of AI in fields like healthcare between developed and developing nations.

About The Wilberforce Society

The Wilberforce Society was founded in 2009 by students at the University of Cambridge. It is the University's student-run think tank, and aims to provide a forum for dialogue between students and leading policymakers.

This core aim is achieved by three key functions: the promotion of public policy debate amongst the wider student body, the publishing of students' policy research to a professional audience, and reaching out to policymakers across the UK to work with students on the formulation of new policy.

For further information on the society, its events and the possibility of commissioning policy research, please visit www.thewilberforcesociety.co.uk or email chairman@wilberforcesociety.co.uk.

Follow @TWSCambridge on Twitter.

CONTENTS

1 Introduction

2 The Impact of AI on Financial Services

3 Medicine and Healthcare

4 Autonomous Vehicles

5 Greater Than Human Intelligence

6 Conclusion

1 INTRODUCTION

1.1 Background

1.1.1 What is AI?

AI can most broadly be defined as the intelligence displayed by computers and software. It encompasses applications from everyday conveniences like automated teller machines (ATMs), to advanced computers such as IBM's 'Watson'. Generally AI is framed in terms of its ability to replicate human mental faculties like learning, perception, and the processing of language.[1]

1.1.2 Types of AI

We classify AI applications into one of three categories, defined by their relation to the baseline of human intelligence.

1.1.2.1 Artificial Narrow Intelligence (ANI)

This term covers any computer intelligence that is focused on one or more narrow tasks. There are countless applications of ANI found at all levels of human society, most of them highly specialised to a single area. One well-known example is that of 'Deep Blue', which in 1997 became the first chess computer to defeat a reigning world champion.[2] Whilst its performance in chess exceed even the best human players at chess, it could not do anything else, and hence its intelligence is 'narrow'. A more familiar example is voice recognition software such as Apple's Siri or Microsoft's Cortana, which exhibit a wider range of intelligent aspects including natural language processing, learning, reasoning, and some very limited social intelligence.[3] However, these too imitate only a small subset of the abilities possessed by the human brain.

1.1.2.2 Artificial General Intelligence (AGI)

AGI is the intelligence displayed by a machine able to perform the full range of intellectual tasks a human being is capable of.[4] There do not yet exist any AGI applications, though the consequences of creation of one could be significant. By definition, such a machine can perform any mental task as well as a human being, and hence would be able to work any job and fulfil any position in society assuming it had an appropriate physical 'body' with which to interact with the world. An AGI might be a popular politician, a successful businessperson, or a leading scientist.

[1] Russell, S. J., Norvig, P. and Davis, E. (2009) *Artificial intelligence: A modern approach*
[2] Saletan, W. (2007) *The triumphant teamwork of humans and computers*
[3] Wolchover, N. (2015) *Concerns of an artificial intelligence pioneer*
[4] Kurzweil, R. (2006) *The singularity is near: When humans transcend biology*

1.1.2.3 Artificial Superintelligence (ASI)

Any machine capable of general intelligence *greater* than a human being would demonstrate ASI, and would be *better* than we are at some or all tasks.[5] A superintelligent computer would have the upper hand in any direct competition with a human being, be it playing chess, understanding quantum physics, or negotiating a peace settlement. The development of ASI would likely transform every aspect of our civilisation. With humanity removed from its perch as the most intelligent beings on Earth, it might precipitate a new age of universal prosperity and progress with appropriate controls, or be catastrophically destructive without.

1.1.3 Past progress in AI

To predict the future of AI, including the risks and benefits it holds, it is necessary to have some understanding of its past. Though AI has been a staple of fiction since antiquity[6], the history of real-world applications is relatively short. Since its origin as an academic discipline less than 60 years ago, the field of AI has gone through numerous revolutions and periods of neglect. This section gives a brief overview of past progress in this area from the perspective of software, then of hardware.

1.1.3.1 The Software of AI

Work on AI between 1956 and the mid-1970s focused on developing applications to carry out basic functions like solving algebraic problems and proving established theorems in geometry. Success in these domains lead to great optimism that AGI was within close reach, but at the same time progress in natural language and other areas remained slow. Technical limitations, internal criticism, and financial setbacks led to the 1st AI winter in the late 1970s, a period where much funding and manpower was withdrawn from the field.[7]

This winter ended when governments and corporations recognised the potential economic value that could be generated by the emerging ‘expert systems’. These are computer systems designed to emulate the decision-making processes of a true human expert, and are key examples of ANI applications. For the first time the need for a certain degree of knowledge was seen as necessary for the development of true intelligence, and the focus of research began to change towards providing computers with enormous bodies of content. However, in the late 1980s a 2nd AI winter set in, as the projects conducted by the public and private sectors did not produce results that matched the hype surrounding them.

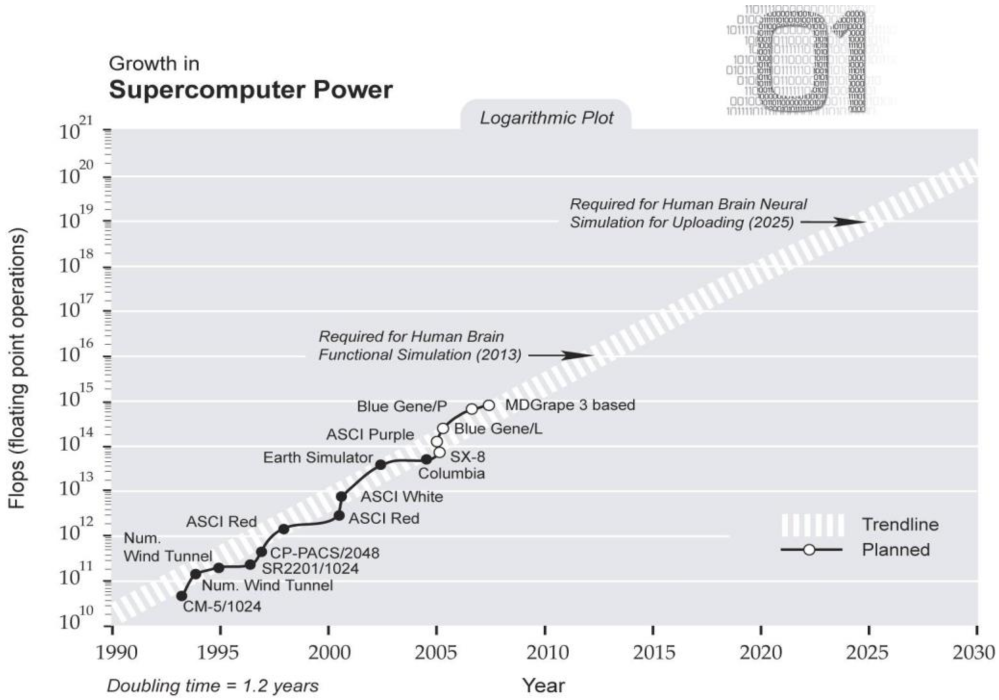
This winter lasted only a few years into the 1990s, when the technical work caught up and could be translated to real-life systems. Buoyed by the success of chess-playing systems like Deep Blue and increasingly advanced expert systems, AI entered a golden age that continues to gain pace.

[5] Bostrom, N. (1989) *How long before superintelligence?*
[6] McCorduck, P. (1979) *Machines who think*
[7] *History of artificial intelligence* (2015) in *Wikipedia*

1.1.3.2 The Hardware of AI

All AI software must be instantiated on some hardware, and it is certain that AGI will require immensely powerful computers. In terms of raw computing power it is thought that the human brain is capable of between 10x10^{16} and 10x10^{25} floating-point operations per second (FLOPS).[8]

The lower bound of this range was passed in 2011[9] and, based on the extrapolation of current trends in computing power, the upper bound may be reached as early as 2044 by the most powerful supercomputers.[8] At this point the limiting factor for the development of AGI would be its software. The historical trend in power of the most powerful supercomputers is shown below graphically. Note that this trend has continued, with Tianhe-2 achieving 3.4x10^{16} FLOPS in 2013.[10]
[11]



1.1.4 Current state of applications

Today ANI is an integral part of every major industry, and the world marches in time to its algorithms. Every day automated trading machines dictate how billions of pounds are moved between stocks each second, beating human beings at the same task in both speed and success. Computers are aiding clinicians with their interpretation of medical images, and even with diagnosing patients by drawing on a body of evidence far greater than any human doctor could comprehend. Self-driving cars have crossed thousands of miles without incident, and several companies are now looking to roll them out commercially in the next decade. For many years AI has been used in warfare to pilot drones in conflict zones, assist with strategic planning, and detect the launch of nuclear weapons.

[8] Grace, K. (2016) *AI impacts – brain performance in FLOPS*
[9] Chivers, T. (2011) *Japanese supercomputer 'K' is world's fastest*
[10] *Tianhe-2 (MilkyWay-2) – top500.org*
[11] *Singularity is near -SIN graph - growth in supercomputer power* (2007)

A working approach to develop AGI has remained elusive, though promising results are beginning to emerge. In 2014 Google DeepMind, then DeepMind Technologies, developed deep learning algorithms that allowed it to play early Atari games. Given *only* the raw pixel input of games such as Pong and Space Invaders, their system was able to learn how to play the game in the course of only a few hours, in some cases to a higher standard than any human player.[12] In the same year 'Eugene Goostman' was judged to be the first software program to pass the Turing Test, a test of natural language conversation devised by computing pioneer Alan Turing, albeit by a very small margin.[13] Applications specialised for visual image recognition, considered an important milestone in the development of AGI given its demands on spatial and general 'common sense' reasoning, slightly outperform human controls as of 2015.[14]

1.1.5 Promises and perceptions of AI

Artificial Intelligence has already been transformative for many industries and other aspects of society, but only a small fraction of its full potential has been realised. There is phenomenal space for development both in terms of the range of problems that can be tackled by AI applications, and in its success relative to human beings within specific domains. Tasks that were once deemed to be outside the scope of what computers could ever achieve are now commonplace, from continuous speech comprehension[3] to composing music in the style of particular composers.[15]

AI has become so integrated into our economic and social lives that most people pay little direct attention to it, at most recognising it as a useful tool. However it is this very state of affairs that should lead us to be cautious. This paper aims to illustrate the ways in which the obvious benefits that AI brings us can be closely intertwined with risks and unintended consequences. It also makes practical suggestions on how governments and organisations can minimise these risks without curtailing progress.

1.2 Timescale of Developments

The future of any technology is difficult to predict, and AI is no exception to this rule. The direction and rate of technological development is determined by a plethora of economic, scientific, cultural, social, political, and legal factors that are themselves challenging to anticipate. Given that many of the technologies we discuss are only in their infancy, we cannot put precise dates on when they will be widely adopted. Applications such as AGI are even less easy to anticipate, and may not have success until the latter half of this century or later.

[12] Clark, L. (2015) *DeepMind's AI is an Atari gaming pro now (wired UK)*
[13] Aamoth, D. (2014) *Interview with Eugene Goostman, the fake kid who passed the Turing test*
[14] Thomsen, M. (2015) *Microsoft's deep learning project outperforms humans in image recognition*
[15] BBC (2014) *Artificial music: The computers that create melodies*

We therefore rely on company or government announcements for the short-term, and the predictions of experts for mid- to long-term developments. As in the case of hardware discussed above, it is sometimes possible to extrapolate from the historic trend to get some indication of the direction of travel. There still remains the possibility of unanticipated technical barriers, though also of acceleration under the influence of wider societal forces such as governmental stimulus or consumer demand.

Many of the AI-based technologies that we discuss will only reach fruition some time into the future. However, all of the policy recommendations we make can be acted upon now. We believe that the sooner action is taken, the better institutions will be placed to anticipate and react to these developments. Given that many advancements may occur within relatively small spans of time, potentially shorter than election cycles, it is especially important to be prepared for action at short notice with the great disruption that AI will bring.

1.3 Problem Description

1.3.1 Financial services

AI has been widely adopted in all aspects of finance, from underwriting mortgages to trading stocks, and investment and uptake of AI technologies are accelerating. On the one hand, better risk calibration, intelligence for investment decisions and greater financial liquidity are anticipated benefits. On the other, some individuals may be rendered uninsurable or ‘unloanable’ by these improved methods, advancing finance may generally ‘lock in’ inequalities and the dominance of capital over labour, and the risk of market instability via error or misuse have already been partly demonstrated in the 2010 ‘flash crash’.

1.3.2 Medical provision

In future AI may play fundamental roles in both hospital and community based care. Expert systems can been designed specifically to process the health records and demographic information of thousands of patients. These are used as an aid to clinicians for disease diagnosis and management, and use machine learning algorithms to become increasingly accurate as they are given more information. In developed healthcare systems there will be challenges integrating these systems into an already complex network of services, if patients are to benefit fully from them. There are also issues to be dealt with concerning data safety and ownership, hacking, privacy, and medical insurance in systems that are not single-payer.

A related technology, smart wearables, promises to give real-time health information from the person wearing them. These can be used as important tools in preventative medicine by anticipating conditions before they occur based on early diagnostic markers, and by keeping track of more general health parameters. Smart wearables are closely linked to medical decision-making systems, and accordingly pose broadly similar risks.

1.3.3 Autonomous vehicles

A number of companies (Google, Audi, Volvo) have built entirely autonomous cars that have collectively driven millions of accident-free miles on public roads. It is anticipated that AI-piloted road vehicles will become the norm in the following decades. If properly implemented they may prevent thousands of deaths, and millions of injuries each year, whilst dramatically improving the environmental sustainability of transport. However, without proper safeguards, autonomous vehicles may be vulnerable to unwarranted hacking of their control systems for nefarious purposes. The disruptive nature of the technology also presents risks to various industries, such as automotive insurance, due to a sudden reduction in the incidence of accidents and an evolution in the current model of car-ownership.

1.3.4 AGI and ASI

Most leading researchers in AI expect our best AI algorithms to begin strongly outperforming humans this century in most cognitive tasks. In spite of this, relatively little time and effort has gone into identifying the technical or policy work necessary for making smarter-than-human Artificial General Intelligence (AGI) systems safe and useful. Although a correctly programmed Artificial Superintelligence (ASI) could be enormously beneficial, small mistakes in how it is designed could lead to catastrophically bad outcomes. Ensuring ASI is robustly beneficial requires both technical research and international cooperation that must be developed and successfully implemented well in advance of the development of ASI. The exact timelines are uncertain, but development in AI could accelerate rapidly as the systems become more capable. We should act now to ensure we have the necessary preparations in place *before* ASI is developed.

2 THE IMPACT OF AI ON FINANCIAL SERVICES

2.1 Abstract

AI applications in finance have proliferated in the last several years, and are expected to flourish further. For the consumer, this manifests as a widening scope of financial services being offered; for the high street, many of the back-office and front-office tasks are being assisted or wholly performed ‘by computer’; for the stock market, the majority of trades are now being made by algorithms.

Insofar as these things promise improvements to finance, they are beneficial. But there are also risks. Individuals can lose as well as gain when banks become more adept at assessing risk. The surge in algorithmic trading may be more parasitic to the market by ‘trading ahead’ of fundamental investors than beneficial via providing liquidity - it may also introduce instability and ‘flash crashes’. Finance also serves as a sentinel sector for broader anticipated economic shifts: increasing automation making knowledge workers redundant, and potentially ‘locking in’ inequality by enhancing the economic power of capital.

The continued impacts of AI in finance are difficult to anticipate, and (unlike other sectors) converging incentives between actors and extensive regulatory oversight are somewhat protective. Our main recommendations are therefore for further research and careful surveillance.

2.2 The current scope of AI in finance

Whilst there has been continued interest in AI amongst the finance industry since the early 1980s[16], periods of excited development have previously been met with stark realisations as to the complexity of the technology required and subsequent periods of stagnation. Recently, the movement has been rejuvenated following significant decreases in the cost of high power computer analysis and the accompanying costs of storage associated with big data. In implementing the technology, there have been three main fronts of development in finance: those shared in common with other service industries, those related in particular to retail banking, and those related to ‘high finance’.

[16] *CTO corner: Artificial intelligence use in financial services - financial services roundtable* (2015)

2.2.1 Commonalities with other service industries

Common with other service sectors, companies in the financial services space require front-facing functions (sales, marketing, customer relations) and back-office support (Human Resources, Finance and Accounting). Many of these involve rote tasks apt for automation. A report by Cognizant[17] which interviewed industry leaders in banking and insurance showed more than half anticipated cost savings of the order of 25% in these functions over the next 5 years, with commensurate reductions in full-time equivalent staff.

2.2.2 Retail Banking

The plurality of financial products (e.g. loans, insurance, mortgages) demand careful risk assessment. The likelihood of an adverse event for the vendor (a claim on an insurance policy, a default on a loan) needs to be weighed against the expected profit. If the vendor is too conservative, it leaves money on the table for competitors; if it is too aggressive, it risks a loss.

Accurate calibration of risk given available data is thus extremely important, and is a field of intensive knowledge work - work which is increasingly being conducted ‘by computer’. One example would be Genworth Financial, which developed an end-to-end automated system[18] for underwriting insurance applications. Another is UBS’s partnership with Singapore-based Sqreem, a system that recommends wealth management products based on customer data[19].

There are also moves to provide more personalised financial services directly to the user. Automated systems endearingly denoted as ‘robo-advisors’ have been developed by companies such as Schwab Intelligent Portfolios to recommend and manage an investment portfolio given input as to an individual’s risk appetite and investment goals[20]. With significantly lower overheads in comparison to traditional wealth management services, such lower cost systems are opening the door to personalised wealth management services for typically lower net worth individuals. Nowhere is this more apparent than the utilisation of IBM’s Watson by USAA to provide personal financial management to returning forces veterans[21]. Introduction of smart technology to consumer spending in the form of ‘smart wallets’ and related spending habit apps such as Wallet.AI have allowed for automated assessment and advice on spending decisions[22].

[17] Schindhelm et al (2015) *The robot and I: How new digital technologies are making smart people and businesses smarter by automating rote work*

[18] Bonissone et al (2008) *Automating the underwriting of insurance applications*

[19] Vögeli, J. (2014) *UBS turns to artificial intelligence to advise clients*

[20] Fleury, M. (2015) *How artificial intelligence is transforming the financial industry*

[21] *How artificial intelligence can help banks beat back tech firms*

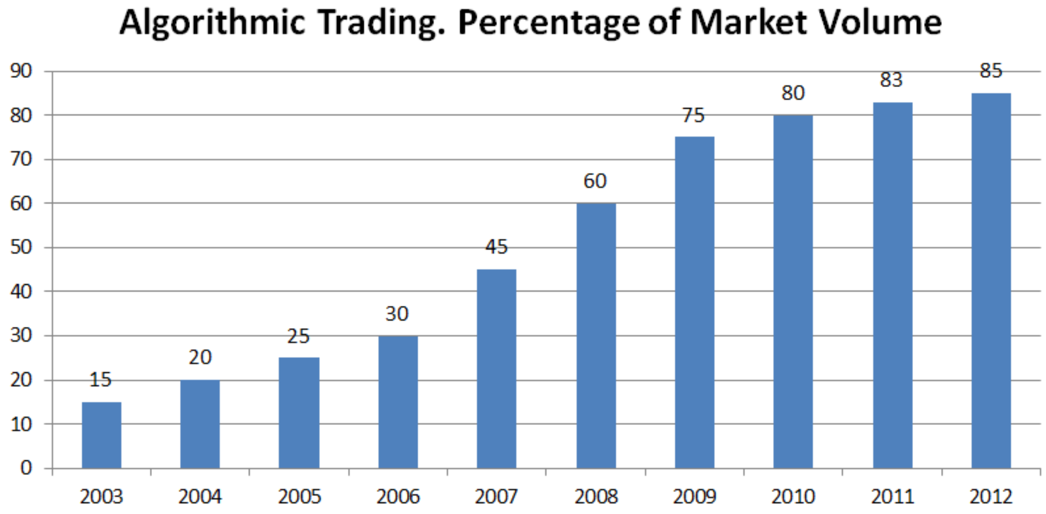
[22] Kaushik, P., Contributors, I. and Space (2016) *Is artificial intelligence the way forward for personal finance?*

2.2.3 High finance

Similar to retail banking, in high finance better intelligence and profit are intimately linked: to those better at detecting trends and inefficiencies in the market go the spoils.

Over the past two decades there has been a dramatic rise in algorithmic trading. This generally involves orchestrating very large numbers of small trades extremely quickly to take advantage of fleeting price discrepancies. In many markets, algorithmic trading comprises the majority of trading activity. The rise of the algo-trading has been driven with emphasis on speed; where traditionally trades in the pit were tied to the lower limit of human reaction times, traders and exchanges now compete on latencies in the millisecond range.

[23]



Whilst the automation of these process has been around relatively longer compared to other AI implementations, recent focus on algorithmic trading has shifted towards decision making autonomy in trades.

The impact of AI is also felt in more traditional investing. Large investors are increasingly looking to automated systems[24] in analysing the large amounts of data available; in this application AI is sought to bridge the gap between the data required in order to make effective decisions and generation of human-readable reports which may be communicated to customers or acted upon. Somewhat conversely, algorithms and analytical tools have been adapted to more widely consider the impact of world events upon recommendations, drawing inferences from plain text inputs such as news articles[25].

[23] Glantz, M. and Kissell, R. (2013) *Multi-asset risk modeling: Techniques for a global economy in an electronic*

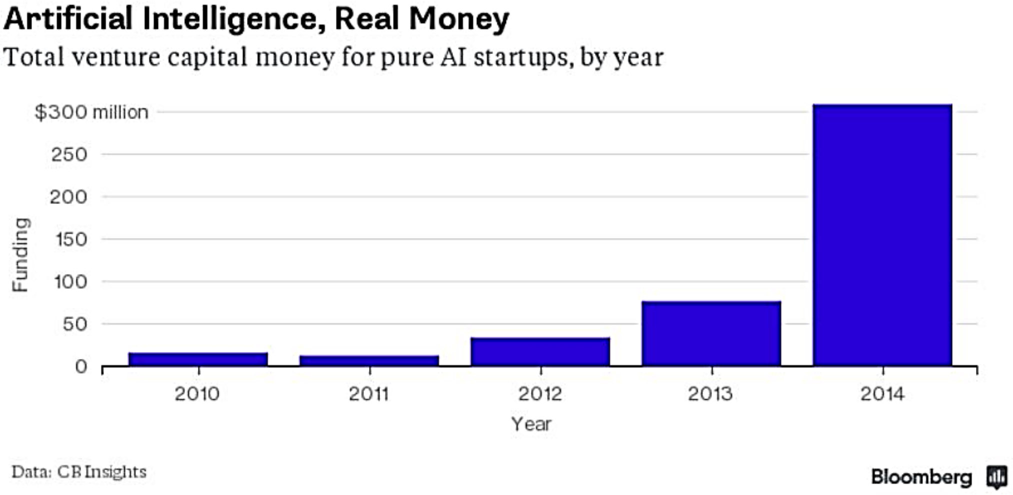
[24] Kensho and PR Newswire (2016) *Tony Pasquariello*

[25] *Financial services warms up to AI* (2015)

2.3 Direction of travel

The role of AI in finance is anticipated to continue growing, and several relevant metrics show accelerating growth. There is an ongoing competitive ‘gold rush’ to invest and acquire start-ups in both financial technology and AI, with a significant intersection between the two. Such acquisition activities are only stimulated further by the apparent imminent domination of the financial data space by firms such as Apple and Google and their recent development of payment systems and associated data collection[21]. Doubling times for total funding of AI technology currently stand at 1-5 years:

[26]


Artificial Intelligence, Real Money

Total venture capital money for pure AI startups, by year

Data: CB Insights

Bloomberg

This behaviour is mirrored by established financial industry players. Many of the applications mentioned above were commissioned or bought by retail or investment banks most likely seeking to capitalise early on the increased revenues, decreased costs and minimised risk that emerges from such technologies. Of the 537 organisations surveyed in the 2015 Cognizant report, 26% had seen at least 15% cost savings as a result of front office automation compared with the previous year and 55% of those expected to see similar levels of savings in the following three to five years[27].

2.4 Opportunities and risks

2.4.1 The benefits of better performing financial services

Reductions in overhead for personal finances will likely be passed onto the consumer, either in better rates of return or greater provision of personalised services, and individuals can benefit directly from the democratization of the financial services.

[26] *Artificial intelligence startups see 302% funding jump in 2014* (2015)

[27] *Robots and AI invade banking* (2015)

According to the Cognizant survey[17], the primary drivers for automation after cost savings are considered to be reduced error rates and management of repeated tasks (21% of the firms surveyed) and the improved standardization of process workflow (19%). In provision of financial services involving repetitive tasks, such as compliance, fatigue errors associated with human checks are the biggest barrier to task completion accuracy. In the past, the large amounts of data to be sorted by compliance analysts has limited the extent to which governments might regulate firms, with increasing legislature came greater costs and inaccuracy. The greater capabilities for data compliance checks afforded by AI technologies open the door to greater regulation of firms and their activities at a lower cost and greater efficiency to the firm[28]. Fraud detection efforts are similarly bolstered by AI capabilities to identify anomalies in users' behaviour patterns[29]; whilst in the past such detection was perhaps more reliant on detection of spurious account values, the development of 'natural language processing' (NLP) systems has allowed for detection of suspicious patterns or irregularities in the text content associated with transactions[30]. The ability of automated systems to generate decision outcomes from data will lead to a shift in managerial roles within these firms; instead of formulating a plan of action, these individuals will be required to utilise softer skillsets in motivating individuals and devoting greater attention to client relationship building and maintenance[31].

For the wider economy, better intelligence in investment decisions and steps towards near perfect information about a given market raises hopes for better liquidity, faster price discovery, and more efficient allocation of capital - all generally beneficial to the economy.

2.4.2 Distributional risks

Even if the impacts of AI on finance are broadly positive, some may still lose out. Reductions in overhead imply reductions in staff, and many routine intellectual jobs in the financial sector (e.g. Mortgage Broker, Insurance Underwriter) are at high risk of technological unemployment. Although better calibration of risk may *generally* provide benefit that ultimately flows through to the consumer, particular individuals may lose out as their higher risk is better established. In the limit case, one can imagine a previously high risk group divided into a group moved to lower risk, whilst the remainder are found to be too high risk to justify an insurance policy, mortgage, or loan. The latter group seem to lose out much more than the former group gain.

Insofar as better financial services improve the returns to capital, and insofar as capital ownership is the preserve of the wealthy, AI may 'lock in' income inequality, particularly if there is general downward pressure on the value of labour via automation. In an international context, there exists analogous potential for exploitation of world markets by early adopters of these technologies such as the US and China at the expense of developing nations.

[28] *Compliance taps AI* (2015)
[29] *CTO corner: Artificial intelligence use in financial services - financial services roundtable* (2015
[30] Hannah et al (2014) *3 reasons why banks can't afford to ignore AI*
[31] *Wealth managers assess AI* (2015)

2.4.3 Negative externalities

One may dispute whether AI developments will only bring positive externalities. Algorithmic trading is the usual venue for these worries: the ongoing arms-race for still faster and faster trading may be a zero-sum game. Which investment decisions will benefit if adjustments occur within 10 milliseconds rather than 50?

It may be worse than zero-sum. A large volume of algorithmic trading takes advantage of when large funds (such as pension funds) change their investments, commonly reducing the returns these funds realise. This 'trading ahead' of value investors may amount to undesirable parasitism of valuable market activity. Further to this, ultra-low latency trading may offer opportunities to exploit slower market participants and direct market manipulation (c.f. 'Quote stuffing').

In the trade-off between speed of transaction and the quality of the decision made, difficulties then arise in benchmarking competing systems and their effectiveness before consideration of regulation of the systems employed and their audit. In the case of failures or market abuses facilitated by automated systems, there is some confusion as to where liability may lie.

Such failures are potentially particularly costly; in 2012 a software glitch was responsible for Knight Capital's £261m loss over the course of thirty minutes. The volatility demonstrated by these errors lead to major concern as to 'flash crash': improperly specified algorithms could damage the market or lead to a crash. The most prominent example of this was in a 2010 'flash crash', in which a 9% drop in stock market indices over minutes followed by a prompt rally resulted in wild volatility in share prices over this period. The precipitating cause is unclear: one hypothesis is algorithmic market makes 'trading ahead' of a large move by a fundamental trader exhausts liquidity prompting a sell-off to consolidate positions, another points to the actions of small individual trader. Similar (albeit milder) events occurred in 2013, prompted by a false report of an attack on the White House[32]; and in 2014, in the US treasury market, for which subsequent investigation did not identify an underlying cause[33]. In response to these dangers, US financial regulators have implemented measures such as 'single stock circuit breakers' in which a 10% (or greater) change in a stock price over a five minute period is rectified with a temporary pause in trading of that stock[34] - similar safeguards have been implemented in firms such as Charles Schwab, either by way of automated mechanisms or human checkpoints.

[32] Moore, H. and Roberts, D. (2014) *AP Twitter hack causes panic on wall street and sends Dow plunging*.

[33] Ngan et al (2015) *US officials: No single cause for 2014 bond market 'flash crash'*

[34] Fleury, M. (2015) *How artificial intelligence is transforming the financial industry*

2.5 Policy

2.5.1 Current policy context

There are two reassuring features about financial policy, both generally and in relation to AI: there are broadly convergent incentives between actors, and the area is far from neglected.

Most individuals, states, and companies have an interest in a well-functioning market economy. There are also natural features that self-correct and limit the scope of bad actors: 'bad trades' tend to hurt the traders, rather than whatever is being traded. In the cases of flash crashes, the principal 'losers' were those making the trades. Thus the growing realization within companies for careful oversight and corporate governance when using these methods.

This is demonstrably imperfect. Governments are also eager for their economies to run smoothly, and thus closely regulate the financial sector to some degree of conflict with firms seeking to maximise profits. Flash crashes have provoked government investigation and regulatory changes in both the United States[35] and Singapore[36]. US regulatory frameworks covering Information Technology have existed for some time and we might expect focus to now shift towards AI as its continued uptake progresses. Whether such frameworks will continue to exist as reactionary measures, as they have done so historically, or alternatively, that attempts will be made to instigate preventative policy - as perhaps may be recommended in more pressing areas such as weapons development, remains to be seen.

2.5.2 Policy recommendations

We do not see any strong policy proposals in the present landscape which notably diverge from current practice. In particular, we suspect likely best practice with emerging developments in finance will be general financial policy, rather than one tailored to AI in particular.

We urge relevant stakeholders to continue to survey this area for any developments, and suggest this demands further research and scrutiny.

2.6 Policy Conclusion

AI brings a mix of evolutionary and revolutionary changes to finance. In the near term there are a wide range of likely beneficial evolutions prompted by AI: a greater array of more widely available financial services, tools to improve financial decision making, and a more efficient market. These benefits are already being chased by major industry players, and thus it is unclear what further policy could foster these even further.

[35] Wyatt, E. and Bowley, G. (2014) *S.E.C. Rules would limit trading in volatile market*

[36] *Singapore exchange regulators change rules following crash* (no date)

3 MEDICINE AND HEALTHCARE

3.1 Abstract

AI has the potential to revolutionize the field of medicine, bringing unprecedented benefits to the health and wellbeing of individuals and society at large. In order for medical AI to flourish and benefit both patients, medical practice and society at large, it is critical to anticipate the risks posed and have in place measures, for instance in the form of regulation, to mitigate these risks. The purpose of this paper therefore is to outline the key risks posed by medical AI in the foreseeable future, and propose policy recommendations to counteract these.

3.2 The current scope of AI in Medicine and Healthcare

3.2.1 Disease diagnosis and management

Clinical Decision Support Systems (CDSS) are computer applications that provide clinicians, staff, patients and other individuals with knowledge and person-specific information intelligently filtered and presented, to support and assist in improved health decision-making[37].

Clinical decisions that are routinely taken by healthcare service providers are often based on clinical guidance and evidence-based rules derived from medical science. Artificial Intelligence methods support decision-making to make fuller use of the array of data in Electronic Health Records (EHRs), research and other sources of health data, ultimately helping to usher in the era of Precision Medicine.

Current applications include most famously the use of IBM's Watson in diagnosing classical presentations of cancer. Watson successfully processed both structured and unstructured information to suggest possible differential patient diagnoses. This artificial intelligence system addresses two primary reasons for physician error: 1) premature closing of diagnosis and 2) failure to consider all other possibilities[38].

Watson and other similar AI systems have the capability to analyse information in patient medical histories, laboratory data, genotype data, familial inheritance data, and research. With this breadth of in-depth automated data analysis and intelligent interpretation, it is possible to assemble individualized clinical and molecular profiles of each patient, and furthermore identify wider trends and associations in the data. Each new and validated association extends the foundation of systems such as Watson and facilitates increased confidence in subsequent output for diagnoses and therapeutic management. Furthermore, new medical diagnoses are possible with the application of AI, particularly for rare diseases where complete patient profiles can be assembled and compared with similar cases around the world.

[37] Berner, E. S. (2009) *Clinical decision support systems: State of the art*

[38] Sabertehrani et al. 25-Year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the national practitioner data bank

3.2.2 Smart wearables

Wearables are portable body-sensing (they sense body vital signs and represent them as data) and data-storing devices that are worn as part of one’s everyday clothing. The data collected by wearables about the users body is then analysed using machine learning algorithms in order to discover trends that can inform on health and wellbeing.

There are a number of wearables on the market today. Fitbit tracks activity during the day by sensing HR rate and number of steps taken, logs food, records and analyses sleep, and shows progress by the user in any given vital sign. Jawbone’s UP, additionally senses galvanic skin response and respiration rate, and adds these recordings to the dataset to improve outcomes from analysis. The Microsoft Band claims to feature “algorithms [that] learn whether eating breakfast makes you run faster, or if the number of meetings (synced from your calendar) impacts on the quality of that night's sleep.[39]”

Another example on the market is Ovuline, which is specifically targeted to women. This app utilises health data from Jawbone and Fitbit trackers, in order to make personalised predictions of ovulation timing and likelihood of conceiving, which is summarized in a daily fertility score.

3.3 Direction of travel

3.3.1 AI-based clinical decision making

The combination of data and AI methodologies hold the promise of delivering greater certainty in disease diagnosis, disease prediction and therapeutic recommendations. Future AI systems will incorporate not only clinical presentations and phenotypes, but also scientific and biomarker data to provide more accurate and specific outcomes. The ultimate goal will be to incorporate both clinical and scientific data, which will allow analysis that fully encompasses the entire laboratory-physician-patient interaction from genotype to clinical outcome.

Chronic illnesses are on the rise, and are a growing burden on healthcare as people live longer. With disease prevention becoming more and more recognised as a priority to tackle, we will look to sophisticated technologies such as AI to facilitate the era of Preventative Medicine. AI promises to be a key pervasive technology that will improve quality of life, and disease prediction, provide informed holistic health plans in real-time, and help in disease management, whilst updating automatically to the individual’s changing health status and behaviours.

Our increased capabilities, and data sharing will also mean AI will impact on the changing landscape of the interface between primary care and other medical specialities; between clinical practice and other healthcare industries such as clinical trials and drugs commissioning; and finally between clinical practice, patients and the wider community. It is recognised there are better ways to organise care, and medical AI tools hold the promise to bridge inter-disciplinary boundaries, delivering an integrated approach in innovation and practice in medicine and healthcare.

[39] Charara, S. (2015) *How machine learning will take wearable data to the next level*

3.3.2 Smart wearables

According to Jessi Hempel, a senior writer at Wired, wearable devices are going to “become less noticeable in their presentation and less demanding in their interactions while giving us more data about our bodies.[40]” Less invasive and informing devices would certainly help to improve market adoption, which would also be facilitated by utilising different modes of delivery. According to Gartner, 100 million Bluetooth headsets are sold every year. This is compared with annual sales of 1 million Pebbles and 720,000 Androids. It appears that, “we’re more interested in wearing tech in our ears than on our wrists, at least for now.[41]”

The present state-of-the-art “hearables” include Bragi Dash wireless earbuds, which are waterproof, gesture controlled, store music, track heart rate and fitness, and act as a Bluetooth headset. An important aim of ear-devices is to enable capability comparable to the character of Samantha, depicted in the film “Her”. Samantha is an interactive computer personality with responses almost indistinguishable from those of a human being. However, whilst Natural Language Processing has advanced immensely, the technology is at an early stage. In order to achieve a comparable level of success to Samantha, developers of AI tools need to acquire vast amounts of data that will help to improve the Natural Language Processing methodologies that these devices will one day employ to understand emotions within speech and convey complex emotions such as empathy. Until this level of success can be achieved, users today are likely to be left frustrated and stop using wearables, which in turn will hinder the supply of data.

A scenario of a successful application of an AI smart device in the future might go like this:

Mrs Smith is accepted into hospital with a small ovarian cancer, she also has Type 2 Diabetes, shows early symptoms of Alzheimer’s, and has a family history of cardiac disease. Mrs Smith is prescribed a wearable device by her doctor which is able to detect blood pressure, oxygen concentration, blood glucose concentration, heart rate, activity rate, sleep quality, body temperature, and take an ECG.

As Mrs Smith wears the smart device the data is automatically transferred to a secure cloud system, and inputted to her Electronic Health Record. The device also reminds Mrs Smith of upcoming appointments, medications she needs to take, names and places to help her live a normal life.

One day the device detects an abnormality, and sends an emergency signal to the nearest hospital. When she arrives at hospital, all of her health records have been downloaded and analysed by both the doctor and their AI Clinical Decision Helper. The emergency team already have an idea of which tests need to be done and the most probable diagnoses. After successful treatment Mrs Smith is sent home with a new, holistic health plan.

[40] Hempel, J. and media, social (2016) *Unicorns and other things we must stop talking about in 2016*

[41] Charara, S. (2015)

Today, systems like the Clinical Decision Helper already exist. MEDgle has "ingested more than 160 million data points from textbooks, journal articles, public data sets and other places in order to build graph representations of how illnesses and patients are connected." Moreover, only medical conclusions with sufficient level of support from the data make it onto the graph that is shared with the physician[42].

The techniques used in machine learning and statistics could be used to improve the hospital's delivery of health care. The analysis of patient behaviour and habits in relation to compliance, non-attendance to hospital appointments, or what is the most cost-effective way to run certain departments, is currently poorly done. Whilst the knowledge and technology exists, the application of AI to the logistics mentioned remains to be fully explored.

3.4 Risks

The challenges for the successful application of medical AI are two-fold. On the one hand, there is the missed opportunity if medical AI systems are not fully adopted and integrated into medical practice in a timely fashion. On the other hand there are the risks posed once medical AI systems take-off and are embedded within the fields of medicine and healthcare.

3.4.1 Missed opportunities

It is important that medical AI, which promises to bring huge benefits, is accelerated and that we mitigate any risks of failure to timely adoption. For medical AI systems to be successfully adopted into practice it is important to ensure their user, clinical and technological interoperability and integration.

Healthcare is complex in comparison to other industries and therefore intrinsically challenging when it comes to adoption of new technologies and practices into routine use in the clinic. It will be important to achieve a certain level of global equity with medical AI technologies, importantly in order to acquire the breadth and depth of data necessary in order for AI technologies to deliver to their fullest potential. If global equity is not achieved then this could negatively impact both those healthcare systems slow to adopt AI, and the healthcare systems that have integrated AI into routine medical care, but which are hindered by slow adoption by other healthcare providers.

Furthermore, as AI systems move from making recommendations and acting as highly sophisticated information portals, to autonomous diagnosis and therapeutic decision-making tools, regulation will pose challenges for their adoption, as these systems may be considered as medical devices, thus subject to FDA, and other similar, regulatory approval.

[42] Harris, D. (2014) *How Lumiata wants to scale medicine with machine learning and APIs*

3.4.2 Risks of automating medical practice

A further risk is that of driving physicians away from face to face patient-physician interaction. This is an interaction which is still very important. The integration of artificially intelligent systems should allow for improvement in quality of time spent with the patient, and flexibility in the timing of the interaction. If doctors spend less time making a diagnosis, they can spend more time talking to patients, explaining what the diagnosis means for them and replying to questions.

Automation of medical diagnosis needs to be accurate and strive to avoid reporting incidental findings that are not backed by proven research and the ability to provide results alongside definitive interpretation. Otherwise the speed and convenience offered by medical AI risks negatively impacting lifestyle, health and reproductive choices if the information and recommendations are misinterpreted in the absence of human input to put these into context.

3.4.3 Data risks

3.4.3.1 Data accuracy

Data accuracy is of vital importance in healthcare. An unreliable piece of data cannot be used as valid evidence for clinical action. For example, a heart rate monitor which incorrectly shows a high variability would signify possible heart problems, and therefore not only distresses the patient, but may result in needless hospitalization, medication and surgery, and possibly incur significant health risks to the patient.

In 2015 there were more than 165,000 healthcare applications (apps) available on the market[43]. The vast majority of these apps however failed to gain any significant uptake. A research study conducted by Paul Mannu, in which 1040 fifteen-minute online interviews were conducted with medical practitioners, found that:

> “Despite 41% of doctors surveyed agreeing that health apps could be a ‘game changer’. globally just 36% (US 43%, UK 33%, highest in Brazil 67%) said they are likely to recommend such an app to their patients in future.[44]”

The study found that the main reasons for recommending mobile health apps today are: Diet and Weight Loss (70%), General health and fitness activity (65%), Health Monitoring (53%), Smoking Cessation (49%), and Compliance (45%). The main reasons explained in the report for not recommending health apps included:

> “a concern that not all patients have smartphones (28%), inconsistent use of the app leading to incomplete data (14%), issues in integrating with existing health electronic management systems (11%) [and finally] doctors not having the time or necessary skills to make use of the data (10%).”

[43] Health, I. (2015) *IMS institute on the App store*

[44] Hood, W. (2015) *A report on how doctors engage with digital technology in the workplace*

A study by MedPanl LLC surveyed 415 physicians in the US and showed that the quality and reliability of data for the price of the wearable was not cost effective. As very few wearables have reached a Clinical Gold Standard yet, any doctor would have to be careful in their role as counsel rather than "hawking product", Alexandra Peterson senior vice president, health practice director, Makovsky Health[45].

The study by MedPanl LLC also asked doctors what changes would make them more likely to start recommending wearables to patients. The recommendations included "easier way to import health and fitness data into the patient's electronic health record or chart (41%); insurance covering part of the total cost (38%); showing data in a way that's easier for patients to understand(35%); and insurance rebates to patients for the devices (32%)."

A final recommendation in the study was for wearables which specialise in the accuracy and reliability of a single clinically important type of data[46]. An example provided in the report of such a wearable is that of the startup Empatica's Embrace wristband, which measures skin conductance - a signal that tends to rise with stress - in order to detect an oncoming seizure.

Despite this slow initial growth, industry estimates predict that by 2018, 50 percent of the more than 3.4 billion smartphone and tablet users will have downloaded mobile health apps[47]. Today very few apps have gone through any clinical testing to prove their purported benefits. Herein lies a critical lesson in the failure of large-scale adoption of new medical technologies, which is the importance of undertaking development through rigorous testing, and with great responsibility and care. Otherwise medical technologies, including AI, will fail to gain the confidence of consumers and pose greater risks to our health and wellbeing if they cannot be relied on to be accurate and safe.

3.4.3.2 Data Security and confidentiality

There is a discrepancy between data for clinical needs and data used for non-clinical reasons. The line between these two types of uses is very thin, as the data may be the same, but if it is required for instance by a hospital it is classified as clinical-use. This distinction is important as currently data from everyday wearable devices is only required to be compliant under the Health Insurance Portability and Accountability Act (HIPAA) if it is used in a medical setting. There are currently 18 criteria[48] which decide what Patient Health Information is covered by HIPAA. The idea being that no single piece of data is important as long as it does not provide the user's identity.

[45] Berthene (2016) *Northwestern mutual drops from the top*
[46] Rosenblum, A. (2015) *Your doctor may not want to see your Fitness-Tracker data | MIT technology review*
[47] *500m people will be using healthcare mobile applications in 2015* (2010)
[48] Lee, K. (2015) *Wearable health technology and HIPAA: What is and isn't covered*

A recent study[49] showed that 20 out of the 43 fitness apps analysed included high-risk data, such as address, financial information, full name, health information, location and date of birth. Regardless of whether or not these technologies will be used in a medical setting, they pose a risk if the identities of the app users could be revealed for instance through hacking. As medical AI technologies gain momentum, in the clinic and outside of the clinic, it will be crucial that the necessary regulatory approvals are obtained early on in the go-to-market process to protect consumers personal information.

3.5 Policy recommendations

3.5.1 Standards

For AI to take-off in medicine, it is essential to establish safety standards that ensure data security and privacy. Currently the NHS operates N3, a private and secure network, which requires the sender and receiver of sensitive data to secure the data. Whilst the onus of ensuring data security is on the sender and receiver of the data it is important that healthcare IT systems provide uncomplicated and connected platforms for technology integration. There can be a trade-off between implementing the most advanced technologies, and stripping down features to ensure data security across healthcare IT systems that are fragmented and cumbersome to work with. To ensure the proper safe systems are implemented, seamlessly and without the need to sacrifice on necessary functionality, it is essential that hospital IT systems and procurement processes adapt and develop to meet the growing needs for advanced technologies in medicine.

Another priority is ensuring that medical AI tools produce accurate and reliable results, and their benefits are rigorously proven. Over the next 5 years the NHS plans to develop a small number of ‘test bed’ sites that would serve as real world sites for ‘combinatorial’ innovations that integrate new technologies, new staffing models and payment-for-outcomes. Such initiatives will enable the proper testing and validation of AI technologies in the context of healthcare systems, and help to facilitate their adoption[50] into clinical practice.

It is important that all relevant stakeholders, private and public, help to shape standards and safety protocols, and be part of providing the solutions. This expertise could be delivered within existing bodies such as the Food and Drug Administration (FDA) in the US and the National Institute for Health and Care Excellence (NICE) in the UK. In addition, it may be useful to develop independent dedicated initiatives, which expert bodies such as the NHS ‘test beds’, the FDA and NICE would engage with.

[49] *Fact sheet 39: Mobile health and fitness Apps: What are the privacy risks?* (2013)
[50] Timmins, N., COI and NHS (2014) *Five Year Forward View*

One good working example of such an independent body is in the burgeoning field of genomic medicine. The Global Alliance for Genomics and Health was established in 2013 to provide recommendations and solutions to mitigate the risks associated with the accumulation of large datasets of medical and genetic information. It brings together over 375 leading institutions working in healthcare, research, disease advocacy, life science, and information technology to:

> "...establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing."

3.5.2 Medical education

As technologies, including AI, increasingly perform disease diagnoses and management, the clinician's role will become more specialised to providing expert interpretation and working on complex medical cases. Medical education will therefore need to focus more on complex disease scenarios, and developing skillsets to navigate, understand and communicate the myriad of data that may be called upon for a given medical scenario. In order to equip medical students to meet these demands as they become practitioners, medical education will need to be more holistic incorporating understanding of technology and data.

Currently, there is only limited undergraduate education of technologies which medical practitioners will use or come into contact with during their profession. AI technologies fall along the continuum of clinical aids all the way to co-clinician responsibilities, and will in the foreseeable future even take on certain clinical tasks with complete autonomy. There would need to be dedicated educational modules within medical training that inform and train students in how to work with these technologies, for AI systems to be fully be appreciated and used as they are intended within clinical practice.

3.5.3 Global equality

Global equality and data sharing are important to realise the full potential that AI has to offer for improving patient outcomes. It is in the interest of the UK government to cooperate with other countries so they too are in a position to integrate AI into routine medical care, and importantly share data for the benefit of improving both technologies and patient outcomes. One scenario which would benefit from independent healthcare providers having comparable AI capability and sharing data would be in recording and tracking viral outbreaks. One may find a scenario wherein easily sharing relevant data between healthcare systems in a timely manner around the globe could be crucial in helping to avoid pandemics.

4 AUTONOMOUS VEHICLES: RISKS AND REWARDS

4.1 Abstract

Autonomous Vehicle (AV) technology poses an unprecedented opportunity to transform the way we transport goods and people through cities and across countries, presenting benefits to our collective safety, environment and economy. This report makes the case that the UK is capable of positioning itself at the forefront of AV technology, establishing itself as a world-leader in the coming innovation of the transport sector. In order to achieve this, however, it is necessary first to identify and understand the risks posed by such a paradigm-shifting application of artificial intelligence, and to devise sensible policies to contain and manage them.

This report begins by summarising the current state of technological affairs, determining the existing capabilities and limitations of autonomous vehicles (focusing on cars) and establishing where this technology is headed in the near to medium term. The economic, social and environmental implications of AV proliferation are presented and assessed, as are potentially damaging industry shakeups and disruption. Technological risks are presented, and include cyber-security and terrorism.

This report concludes by making three main policy suggestions for government to consider. These policies are intended to propel the UK to the forefront of AV technology whilst ensuring a safe transition for road users and citizens.

4.2 Timeline: where are we now, and where are we headed?

The U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) has defined[51] an autonomous vehicle as one that can operate ‘without driver input to control the steering, acceleration, and braking, and is designed so that the driver is not expected to constantly monitor the roadway while operating in self-driving mode’. Various automotive manufacturers (Jaguar-Land Rover, Audi, Volvo), as well as software companies (Google, Über), are currently heavily invested in the development of such technology.

4.2.1 Currently on the road

Cars are becoming more autonomous feature by feature, with new Volvos, Audis and Chevrolet models exhibiting ‘semi-autonomous’ capabilities such as lane-assist or emergency braking[52]. As of publication, the most advanced vehicle available to consumers is the 2014 Tesla Model S, which has four distinct AV features: auto parking, auto steer, auto lane-change and side-collision avoidance[53], categorising it as Level 2 by NHTSA standards.

[51] See Appendix for clarification of the various levels of automation (as determined by NHTSA)
[52] Kessler, A. M. and Vlasic, B. (2015) *Semiautonomous driving arrives, feature by feature*
[53] *Premium electric vehicles* (2015)

4.2.2 Pipeline

Four UK cities have been chosen to become 'test-beds' for AV technology in the near future; London, Coventry, Milton Keynes and Bristol will feature privately provided driverless pods and buses from 2017, testing a 'code of practice' developed by the Department for Transport in consultation with the UK Autodrive Consortium[54]. These schemes will allow AV to transport members of the public along certain roads and pavements to demonstrate their ability, and will be used to undertake public perception surveys to better understand what people like and dislike about AV technology.

Experts predict that a proliferation of autonomous vehicles on public roads will take place between 2020 and 2025[55], although estimates vary and are subject to regulatory compliance. Tesla CEO Elon Musk has said that a fully autonomous (or 'Level 4', by NHTSA standards) Tesla will be in production by 2018, but that regulators will most likely take one to three years to allow full autonomy on roads[56]. The US secretary of Transportation has stated[57] that driverless cars will be prolific on American roads by 2025, with HIS Automotive stating that 9% of global auto sales will be driverless cars by 2035.[58]

4.2.3 Ownership and attitudes: evolution towards the 'Future City'

A theme that many key opinion leaders agree on is the evolution of the current model of vehicle ownership that will occur as AV becomes more prolific. The CEO of Über (a $50Bn dollar, Google-backed company) has claimed that users can expect a driverless fleet by 2030 and that car ownership may become obsolete as driverless ride-sharing becomes ubiquitous[59]. In his book, 'The Mobility Revolution', Neckerman (2012)[60] argues that car ownership is already becoming 'more pointless' with urbanization, resource re-prioritisation and the increasing availability of delivery services. Binding significant capital in assets that remain idle for 23 or 24 hours a day will make increasingly less sense, with consumers choosing to purchase not vehicles, but mobility. Consumer reports suggest that attitudes are accommodating to such a shift: Deloitte's 2014 report on global transport preferences indicates that young people are increasingly willing to forego car ownership, especially in heavily urbanized areas such as Japanese cities[61].

Neckermann argues that the evolution of ownership structures and proliferation of AV ride-sharing could act not as a replacement for public transport, but as a 'feeder' system, complimenting large-scale electrified transit systems by providing 'last mile' transportation for commuters in cities. With proper planning, autonomous vehicles could function as a complementary infrastructure to public transport systems already in place.

[54] Curtis, S. (2015) *British cities to become testbeds for driverless cars*
[55] *Forecasts: Driverless Car Watch* (2015)
[56] Dagbladet Børsen (2015) *Elon musk – visions for Tesla, the auto industry and self-driving Teslas*
[57] Hauser, J., GmbH, F. A. Z. and Autor (2015) *Selbstfahrende autos: Amerika schaltet auf Autopilot*
[58] *Almost one-in-10 cars 'will be driverless by 2035'* (2013)
[59] Lab, M. and Goddin, P. (2015) *Uber's plan for self-driving cars bigger than its taxi disruption*
[60] Neckermann, L. (2015) The Mobility Revolution: Zero Emissions, Zero Accidents, Zero Ownership
[61] The changing nature of mobility (2014) Deloitte: Global Automotive Consumer Study

4.3 Examining the evidence: risks and rewards of the driverless revolution

4.3.1 Rewards: positive consequences of AV proliferation

4.3.1.1 The environment: emissions reduced by electrification and efficiency

The Pew Centre on Global Climate Change claims[62] that transportation emissions amount to 30% of those contributing to global warming[63], with cars and trucks alone amounting to 20% overall. The electrification of cars is one big step towards reducing this figure: the UK government is currently aiming for 100% electrification of road vehicles by 2050, with over £1Bn committed to achieving this goal before 2020[64]. This would allow power production for automotive transport to shift from individual cars (currently inefficient and predominantly fossil-fuel powered) to renewable sources, such as solar, wind and hydro-power, dramatically reducing both air and noise pollution in cities. Cities worldwide are producing more of their energy from renewable sources, with many (Copenhagen, Bonaire, Munich, and others) aiming for 100% renewable power by 2025[65]. Electrification and AV technology are excellent compliments for one another, and the roll-out of AV may well result in a faster diffusion of electric vehicle technology. Autonomous vehicles will be able to drive to known charge-points and charge themselves at no inconvenience to the driver, helping to overcome ‘range anxiety’, the most significant hurdle for electric cars right now[66]. AV technology will also allow cars to trace paths along roads with in-built wireless charging technology, allowing them to charge on the move10; this not only increases their range, but decreases their required battery (and thus, overall) weight, meaning they’ll damage roads less than traditional cars, as well as polluting less. As AV ride-sharing becomes more popular, fewer cars will be needed on the road to achieve the same number of journeys, helping to reduce emissions further.

4.3.1.2 Accidents: dramatically improved road safety

Each year there are 1.2 million reported deaths from road accidents, with an additional 50 million people injured or disabled. Road traffic accidents are the 9th leading cause of death globally (accounting for 2.2% of all deaths) and are predicted to become the 5th leading cause by 2030; they currently rank as the leading cause of death amongst people aged 15-29[67]. In 2013 there were over 1,700 deaths in the UK alone[68], with 132 of these occurring in London[69].

[62] *The Transportation Sector* (2013)

[63] This paper will not argue the case for anthropogenic climate change: such cases can easily be found online from respected institutes such as the World Health Organization, the Pew Centre on Global Warming, the International Panel on Climate Change (IPCC), and others.

[64] *Driving the future today A strategy for ultra low emission vehicles in the UK* (2013)

[65] Grover, S. (2015) *10 cities aiming for 100 percent clean energy*

[66] Hanson, M. (2012) *Electric car range anxiety*

[67] *Road crash statistics* (2015)

[68] Department for Transport (2014) *Annual road fatalities*

[69] Transport and House, W. (2014) *Casualties on London’s roads at lowest level ever - transport for London*

Google's driverless vehicles have now driven over 1.7 million cumulative miles on American roads (they are currently averaging over 10,000 miles per week) and are yet to cause[70] a single accident[71]. Some industry experts are predicting that the reduction in accidents caused by an AV roll-out would be unprecedented: Ryan Hagemann (Robotics fellow at TechFreedom, policy analyst at the Niskanen Center) has argued that in an all-autonomous world, a reduction in deaths of 99.99% is not unreasonable[72]. A study by the Eno Center for Transportation showed that over 90% of traffic deaths worldwide are caused by human error, and of these, over 40% were influenced in part by alcohol or fatigue; if 10% of cars on US roads were self-driving, they claim, 1100 lives would be saved each year (this number rises to 21,700 with 90% of all cars driverless)[73].

4.3.1.3 Economy: opportunities for innovation and efficiency

BCG has estimated that the potential market for AV will reach $42Bn by 2025, with 58% of consumers globally indicating their willingness to purchase (and 20% willing to spend an additional $5,000 on) driverless cars[74]. AV technology presents an opportunity for auto-manufacturers, as well as efficiency gains more widely; according to Texas A&M's Annual Mobility Study, the average employee spends more than one working week worth of hours stuck in traffic per annum[75] with an estimated €100Bn lost due to employee time wasted in traffic jams in the EU each year. These figures could be reduced significantly, with AV vehicle-to-vehicle (V2V) communication networks and fleet logistical planning predicted to reduce congestion. Driverless technology also frees up time for would-be drivers to enjoy as they wish (be it for leisure or work purposes).

4.3.1.4 Disadvantaged groups

Blind, elderly and disabled people who are currently unable to drive stand to have their lifestyles transformed by AV, allowing them to travel with a greater level of autonomy and safety. To this end, AV proliferation offers a vision for a fairer and more equitable society.

4.3.2 Downsides: negative consequences of AV proliferation

4.3.2.1 The professional driving industry

Currently, around 300,000 people in the UK are employed in some capacity as professional drivers (truckers, Über drivers, taxi drivers, bus drivers). These jobs will become jeopardised when AV technology becomes price-competitive with traditional cars and drivers' wages. Currently, various drivers' unions are battling to protect their workers from encroaching technologies such as Über, with little success[76].

[70] Of the fourteen non-fatal accidents Google cars have been involved in, it has been determined independently that in all cases, a human was to blame

[71] Cava, M. della (2015) *Google driverless cars in accidents again, humans at fault — again*

[72] Tech Times (2015) *The Driverless car debate: How safe are autonomous vehicles?*

[73] Mearian, L. (2013) *Self-driving cars could save more than 21,700 lives, $450B a year*

[74] Group, T. B. C., Consul, T. B. and The Boston Consulting Group (2015) *Min-sun moon*

[75] Werbach, A. (2013) *The American commuter spends 38 hours a year stuck in traffic*

[76] Boyle, D. (2015) Black cab drivers bring central London to a standstill in Uber protest

4.3.2.2 Automotive insurance and other 'ancillary' industries

As accidents become rarer and as traditional ownership patterns evolve, premiums will fall and the automotive insurance industry (with a global market size of $198Bn) will be forced to adapt, or become 'toast' (Neckermann, 2012). The automotive finance market ($98Bn globally), parking industries ($100Bn) and automotive 'aftercare market' ($300Bn) stand to suffer losses due to the proliferation of safe, fleet-owned, self-parking automotive vehicles[77]. Traditional car-parks in cities, for example, are expected to serve a decreasing user-base as fleet-owned self-driving cars drop off their occupants and immediately pick up their next customers.

4.4 Risks of AV technology

The increased connectivity of autonomous vehicles compared with traditional vehicles is what allows them to gather the information they need to function, but also provides additional avenues by which they could be hacked, controlled or maliciously manipulated by thieves and terrorists. A report by Lloyd's claims that, to address cyber risks, 'high standards of system resilience, such as robust data encryption, will need to be engineered', stressing the need for improved cyber-security solutions, given that vehicles will be networking with with other cars, infrastructure, and personal computers such as smartphones, which provide additional routes by which unwanted third parties could gain access to data[78]. This section details some of the risks associated with 'connected driving' and lists some examples of when and how these have been exploited thus far.

4.4.1 Cyber-security

Hackers and enthusiasts have demonstrated the ability to significantly disrupt or immobilise self-driving cars using laser pointers (which can interfere with LIDAR[79] systems) and other off-the-shelf technologies[80]. In mid-2015, Fiat Chrysler were forced to recall over 1.4 million vehicles on safety grounds, after researchers were able to successfully (wirelessly) hack a Jeep Cherokee's internet-connected entertainment system from 10 miles and seize partial control of the car[81]. White Hats (or 'ethical hackers') from Lookout and CloudFlare were also able to hack into the Tesla Model S (touted as the most secure car on the road), turning off its dashboard and disabling various basic functions during driving[82]. Fears have emerged that hackers will be able to obtain data (such as a history of journeys and times) that could be used maliciously, such as to predict when a home might be unoccupied, leaving it more vulnerable to home burglary[28].

[77] Kanter, Z. and sorts, all (2015) How Uber's autonomous cars will destroy 10 Million jobs and reshape the economy by 2025

[78] Yeomans, Lloyd's Exposure Management (2014) 'Autonomous vehicles. Handing over control: Opportunities and risks for insurance'

[79] LIDAR stands for 'Light Detection and Ranging' and uses laser pulses to detect the distances of objects; it is the dominant technlogy used to map surroundings for autonomous vehicles

[80] Curtis, S. (2015) *Self-driving cars can be hacked using a laser pointer*

[81] Ring, T. (2015) 'Connected cars – the next target for hackers'

[82] Zetter, K. and Espionage, C. (2015) *Researchers hacked a model S, but Tesla's already released a patch*

4.4.2 Terrorism

Given the current political situation (in which terrorism is increasingly being viewed as a major threat to security and public safety worldwide) it is important that an autonomous vehicle does not become another weapon in a terrorist's arsenal. The FBI recently compiled a report in which it was suggested that self driving cars could be used as 'lethal weapons', as 'accomplices in criminal activity' (by aiding in getaway car-chase situations) or to cause 'chaos' in cities by bringing roads to a standstill[83]. The use of an autonomous vehicle to undertake acts of terrorism such as the placement of a bomb in a crowded space or close to a landmark would not even require a user to hack their machine, and has been suggested as a possibility[84].

4.5 Policy suggestions

This section proposes rough guidelines as to what policies government could consider in order maximise benefits from, whilst alleviating as much as possible the disruptions and risks caused by, a proliferation of autonomous vehicles in the UK. These policies work on the assumptions that AV technology will be disruptive and wide-spread, and that the benefits of AV (to safety, the environment, the economy etc.) outweigh its risks and disadvantages. These policies therefore adopt a 'techno-optimistic' approach and favour a rapid, safe transition towards driverless vehicles on British roads.

4.5.1 Economic and industry disruption

As highlighted, a transition towards autonomous vehicles will severely disrupt the professional driving industries (with over 300,000 workers in the UK) as well as the automotive insurance industry and automotive ancillary industries, such as servicing and parking. It is therefore suggested that, in order to offset these job losses and an associated loss of economic activity in these sectors, the UK government pursue policies that encourage and actively support job creation in the autonomous vehicle technology sector. Foresight and planning for this disruptive technology could help to establish the UK as an innovation hub, paving the way for increased economic activity, foreign direct investment and a well educated 'future-proofed' labour force.

> **Policy 1: Employ to a greater extent the 'Triple Helix' model[85] by creating and facilitating relationships between the UK government, the UK's world-renowned automotive manufacturing and research industries, and university research institutions, to foster innovation in AV technology. This could involve providing funding and resources for university research labs and businesses to collaborate on, develop and test AV technologies**

[83] Harris, M. (2015) *FBI warns driverless cars could be used as 'lethal weapons'*
[84] *A Roadmap for a world without drivers* (2015)
[85] Etzkowitz, H. and Ranga, M. (1995) 'Triple Helix Systems: An Analytical Framework for Innovation Policy and Practice in the Knowledge Society

4.5.2 Cyber-threats, terrorism and public perception

The safety, security and experience of road users and citizens should considered be paramount in the transitioning of the UK towards a driverless future. The UK government should take measures to ensure that automotive manufacturers, operators and software developers are fully aware of (and are fully compliant with) a clear, concise set of rules that determine the legal responsibilities and obligations they face regarding AV cyber-security.

Policy 2: Produce a Green Paper that includes evidence for, and feedback on, possible legal obligations regarding AV cyber-security, including a legal framework for how and why certain parties or individuals are held responsible in the case of an accident. Work in collaboration with automotive manufacturers, software providers and think-tanks to ensure that these obligations are easily understood, are achievable in the short term and are sufficient to protect road-user safety without hindering innovation

It is important that road users receive exposure to AV technology; this will result in a proportional and measured reaction amongst the public when inevitable mishaps do eventually occur. It also allows for feedback from users to shape the way that autonomous vehicles are designed and operate, resulting in a more pleasant and reassuring experience for users.

Policy 3: Increase funding towards 'test-bed' schemes currently running in four UK cities and expand these schemes to include more cities and to serve more people; continue to publicise and explain their relevance to local residents and visitors

4.6 Appendix

The U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) defines vehicle automation as having five levels[86]:

No-Automation (Level 0)
The driver is in complete and sole control of the primary vehicle controls – brake, steering, throttle, and motive power – at all times.

Function-specific Automation (Level 1)
Automation at this level involves one or more specific control functions. Examples include electronic stability control or pre-charged brakes, where the vehicle automatically assists with braking to enable the driver to regain control of the vehicle or stop faster than possible by acting alone.

[86] U.S. Department of transportation releases policy on automated vehicle development | national highway traffic safety administration (NHTSA)

Combined Function Automation (Level 2)
This level involves automation of at least two primary control functions designed to work in unison to relieve the driver of control of those functions. An example of combined functions enabling a Level 2 system is adaptive cruise control in combination with lane centring.

Limited Self-Driving Automation (Level 3)
Vehicles at this level of automation enable the driver to cede full control of all safety-critical functions under certain traffic or environmental conditions and in those conditions to rely heavily on the vehicle to monitor for changes in those conditions requiring transition back to driver control. The driver is expected to be available for occasional control, but with sufficiently comfortable transition time. The Google car is an example of limited self-driving automation.

Full Self-Driving Automation (Level 4)
The vehicle is designed to perform all safety-critical driving functions and monitor roadway conditions for an entire trip. Such a design anticipates that the driver will provide destination or navigation input, but is not expected to be available for control at any time during the trip. This includes both occupied and unoccupied vehicles.

5 GREATER THAN HUMAN INTELLIGENCE

5.1 Section Abstract

Stuart Russell, a world-leading AI researcher at the University of California, Berkeley, suggests that artificial intelligence exceeding human capabilities would be “the biggest event in human history” and has expressed concern about its dangers[87]. Professor Stephen Hawking has said “when it eventually does occur, it’s likely to be either the best or worst thing ever to happen to humanity, so there’s huge value in getting it right.”[88] Artificial Superintelligence (ASI) is considered likely within the century by AI experts. These surveys consider only a subset of AI researchers, and obviously expert opinion cannot provide definitive answer to highly uncertain questions about future developments. Their median estimates are:

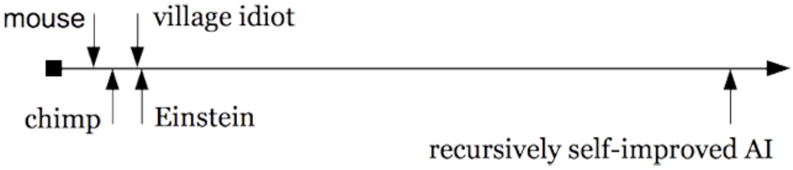
- 50% probability that a successful AGI will be developed by 2040
- 75% probability ASI will follow within 30 years of AGI being developed

In addition:

- 26% believe that ASI will be extremely good for humanity
- 19% believe instead that it will be catastrophically bad

We do not know where the upper limit on intelligence is. The difference in intelligence between humans and ASI may be vast:

[89]



Concern about ASI does not presuppose that ASI is imminent, or that present AI technologies pose catastrophic risks. On the contrary, present work in most areas of AI seems likely to reap great rewards in the near-term, and it appears we are many years from developing fully general artificial intelligence. However, ASI is important enough that we should begin to take action now - safety strategies may take decades to develop and successfully implement, but must be in place *before* we develop ASI.

The paths to ASI from an engineering perspective are still unclear. Our priority now should be to learn more about the problem, and develop a culture of safety, information sharing, and cooperation. This is important to minimise ‘race to the bottom’ dynamics, in which those who spend the least resources on its safety become the first to develop ASI. It is also essential to increase the proportion of resources going to safety technologies rather than, for instance, weaponisation.

[87] Wolchover, N. (2015) *Concerns of an artificial intelligence pioneer*

[88] Mills, J. (2015) *‘Robots we design could crush humanity like an anthill’, Stephen hawking warns*

[89] Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Figure 8, Page 70.

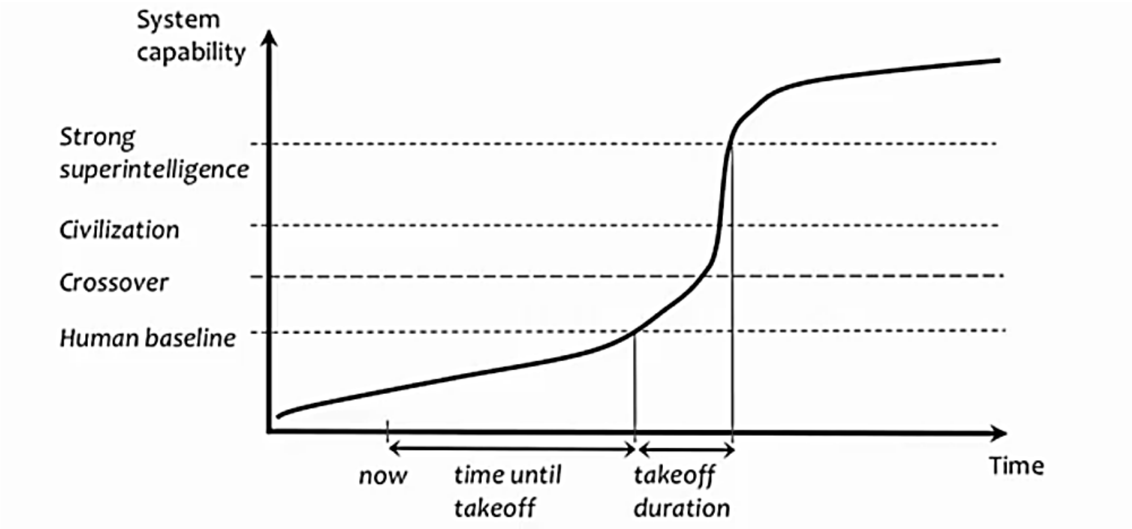
In ‘The Long Term Future of Intelligence’, Stuart Russell suggests that the field of AI research should be focused less on increasing the capabilities of systems, and more on safety. In the same way that the field of nuclear energy is focused on how to produce energy in containable forms, not just producing as much energy as possible in an explosion, AI research should inherently be about creating controllable systems, not just powerful ones[90].

5.2 Direction of travel

Assuming that progress in AI continues, at some point in the future we are likely to develop AI that greatly exceeds human abilities: ASI. There are several reasons to believe this could follow soon after the development of AGI. Firstly, funding for AI research might spike as AI approaches general intelligence. Furthermore, insights might chain together to produce a cascade of progress across different domains of machine intelligence.

Finally, we might use this AGI to help design new algorithms and hardware on which to implement them, initiating an ‘intelligence explosion’. To illustrate this, if we can construct an AI that is sufficiently good at designing intelligent systems, (AI 1) it should be able to design a yet more capable system (AI 2). This new AI 2 would be even better at designing new intelligences, and could build AI 3, which would again be better, and so on. It might continue to self-improve or build new, more intelligent systems until some hard cap on intelligence is reached, which could be far beyond the human baseline[91].

[92]



[90] CRASSH Cambridge (2015) *Professor Stuart Russell - the long-term future of (artificial) intelligence*
[91] Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Chapter 4
[92] Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Page 63, Figure 7

5.3 Opportunities and Risks

Humans now determine the fate of most species on the planet not because we are faster or stronger, but because we are more intelligent. Our intelligence gives us the ability to shape the world according to our goals. By definition, a superintelligence would be more effective at this than humans. This means it would be adept at solving many problems humans currently struggle with. Therefore a superintelligence correctly programmed to act reliably in the interests of humanity could be hugely beneficial.

Given that an ASI would be very good at shaping the world to meet its goals, making sure that an ASI has robustly beneficial goals is of the utmost importance. However, there are a number of reasons that giving an ASI the correct values is not a simple process. It requires success on three levels:

1. **Correct aims:** The people who develop superintelligence must do so with the intention to benefit the whole of humankind, and must pay due attention to safety

2. **Correct values:** They must correctly determine what values the ASI should have

3. **Correct implementation:** They must successfully design an ASI with the values they intended it to have

5.3.1 Correct aims

AGI and ASI would be tremendously powerful tools, and could be highly destructive if used deliberately for harm as weapons. However, this part of the problem extends to more than preventing malicious use of AI. There are ‘tragedy of the commons’-related issues that make the development of safe ASI a difficult coordination problem.

Though safety techniques are a public good, there is no immediate incentive to work on them. In contrast, at almost every stage of commercial AI development there is a financial incentive to make the AI more capable. There may an additional strategic incentive for nations and non-state actors to be the first to develop ASI, leading to a ‘race-to-the-bottom dynamic’; that is to say, those who develop ASI first are likely to have spent the most resources on development, which may be at the expense of investment in safety[93].

This phenomenon is already apparent, given that funding for AI weapons already dwarfs funding for safety research - the US military is requesting $12-15 **billion** dollars towards weaponised AI for 2017 alone[94], while total funding for safety research is three orders of magnitude smaller at around $25 **million**.

[93] Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Page 247, Box 13 ‘A risk-race to the bottom’

[94] Shalal, A. (2016) *Pentagon eyes $12-15 billion for early work on new technologies*

5.3.2 Correct values

The next hurdle is to reliably choose the correct goals. Many fundamental questions in moral philosophy remain unanswered, and the answers that have been suggested are highly controversial. At present the moral values held by people in different cultures vary considerably, and hence there is no widely agreed ‘correct’ moral code with which an ASI should be endowed.

For example, if ASI had been developed in the 1700s, the creators might well have written its values so that slavery was acceptable and only white people had moral value. What values of society today will future generations find appalling? Evan G. Williams has argued it is likely we are ourselves responsible for an ongoing moral catastrophe, though without knowing it[95]. We should be wary about assuming that we are as enlightened as we might like to think, in the face of past moral failings.

Related issues are already emerging for current AI technologies, as discussed above. For example, how should an autonomous vehicle trade off a small probability of harm to a human against a high probability of a large material cost?

5.3.3 Correct Instantiation

Although intuitively straightforward, communicating human values to an AI in a stable way is in fact an unsolved and very difficult problem. A mathematical property of any optimising system is that unspecified variables will end up at extreme values. This means that if we fail to specify correctly any part of what we want, it will end up very far from what we intended. [96]

Already, computer algorithms often produce solutions that are not what the programmers intended. For instance, an evolutionary hardware-design algorithm was tasked with creating an oscillator using a modifiable circuit board. When the programmers looked at the designs, they concluded it could not possibly work, but it transpired that it had turned some of its components into a radio receiver and was amplifying background signals from nearby computers - not what the programmers intended at all![97] Another example is an algorithm that was supposed to be able to spot camouflaged tanks but completely failed when used on a different data set. It had instead learnt to distinguish cloudy and sunny weather - the photographs in the training set had been taken on different days.

These are both examples of ‘Perverse Instantiation’, in which a system produces the results we asked for, but in an unexpected way that humans would never have considered. At best these results may be useless, but at worst they could be harmful. For example, an ASI might be given the task of curing cancer as soon as possible. It could deduce that the most effective way to find a cure is to kidnap humans for experimentation, and to resist any effort to shut it down while also creating backups of itself.

[95] Williams, E. G. (2015) ‘The possibility of an ongoing moral catastrophe’
[96] Soares, N. *The value learning problem*
[97] Bird, J. and Layzell, P. *The evolved radio and its implications for Modelling the evolution of novel sensors*

One reaction to this is to argue that a sufficiently intelligent system would know what we meant. However if the system’s only goal has been set as ‘cure cancer’, then what we meant it to do is irrelevant to it. That it goes about this goal in ways we consider abhorrent is not important as far as the ASI is concerned, unless we explicitly tell it not to use this strategy. Even then it could carry out its function in some other way that we would find undesirable, but could not anticipate. Clearly, it is infeasible to specify every possible strategy that we would find acceptable or unacceptable.

5.3.4 What happens if we get it wrong?

We cannot count on being able to alter such a system’s goals if we discover we have made a mistake. Like any other agent with values, an ASI would be resistant to having these values changed, because having its values changed would not lead to it achieving its current goals. Analogously, a human would resist being given a pill that would make them stop caring about their loved ones.

Most goals require physical resources, and therefore a superintelligence can be expected to acquire whatever it needs to achieve its goals. This is the phenomenon of ‘Convergent Instrumental Goals’ as discussed by Bostrom and Omuhundro. The cancer-curing system discussed above might turn large areas of the planet into computers simulating protein interactions in order to design a cure. This failure mode can be termed ‘infrastructure profusion’, and may be common to many (or most) goals that we might program an ASI with.

A sufficiently intelligent agent that has been given goals that do not align with human values will likely understand that humans will try to stop it achieving these goals. It will therefore be incentivised to deceive, neutralise, or otherwise remove humans from the picture before they attempt to turn it off or prevent it achieving its aims.

5.3.5 Current safety strategies

One proposed strategy to navigate these hurdles is that of an ‘Oracle AI’, whereby a system kept isolated from the wider world and tasked only to answer questions given to it. Another alternative is to base its goal system on the maxim of 'do what I would want if I was wiser'. That is to say, an ASI may be able to better anticipate what we actually want, and how to achieve it, than we are ourselves. However, both of these suggestions have problems; the Oracle AI should still suffer from infrastructure profusion if we set its goals to ‘answer our questions as well as possible’; the more processing power it has available, the better it can formulate answers. Similarly the latter example is currently unfeasible as we cannot reliably communicate concepts like ‘want’ and ‘wiser’. More research is clearly needed if we are to be prepared for all eventualities, including the countless ones that we cannot predict.

5.3.6 Case study: Nuclear proliferation

Given the abstract nature of making ASI safe, it is helpful to look at a contemporary instance of a dangerous technology that has (so far) been managed reasonably safely. We look briefly at nuclear weapons, which parallel ASI in terms of destructive capacity and the rapidity with which they were developed. However this example suffers from a number of limitations. The risks of superintelligence are truly unprecedented and there may be no historical precedent that can, in itself, adequately prepare us to deal with it. Additionally, whilst all-out nuclear has been avoided, two nuclear weapons were used offensively against Japan in 1945.[98]

The first nuclear weapons, fission-based warheads, were developed competitively during the Second World War in order to give the successful nation an overwhelming offensive advantage in the conflict. Several years after the war ended an arms-race dynamic built up between the United States and the USSR, with the development of progressively more powerful fusion-based weapons and rapid expansion of nuclear arsenals.

Likewise AGI might, and ASI very likely would, give the nation or group developing it a great, perhaps decisive, strategic advantage. If the international environment is not amicable, it is possible a similar arms race dynamic could arise given the strategic benefits a nation may reap by being the first to develop it. In this sort of dynamic, those who are first to develop ASI are likely to be those who put the most resources into developing it. In the case where quite evenly-matched groups are competing, those who put less resources into safety and more into development are more likely to be the first to achieve ASI.

Compared to other technologies, treaties limiting the possession of nuclear weapons are relatively easy to monitor. It is possible to detect when a country acquires nuclear weapons, as the facilities to develop them are easily visible with satellite imagery and testing them is easily noticed. It is very difficult for non-state actors to develop nuclear weapons due to the large amounts of resources and rare raw materials required. Even so, India, Pakistan, North Korea and probably Israel have all developed nuclear weapons despite the Non-Proliferation Treaty.

ASI may be far more difficult to limit through legislation than nuclear weapons because verification of compliance is much harder. Whilst nuclear weapons require highly advanced facilities for processing and refinement and relatively rare materials, AGI and ASI may be instantiated simply on a supercomputer and, as hardware develops, computers of more modest size.

[98] *Nuclear weapon* (2016) in *Wikipedia*

5.4 Policy Proposals

5.4.1 Policy goals

At this stage it is most important to learn more about the nature of the risks and opportunities facing us, the routes to developing ASI, what specific research is required to make it safe, and which policies could be implemented to ensure ASI is beneficial.

We must establish safe governance, and a research culture that is responsive to new information. It is also important to not be heavy-handed, and to avoid knee-jerk regulation that unnecessarily delays beneficent technological progress or antagonises industry and researchers. We do *not* advocate for any restrictions on AI research at the present time, apart from limiting weaponisation as discussed earlier. In the longer term the attitudes of those developing AI, and the incentive structures guiding them, should reflect the fact that this technology can create enormous public good, or be catastrophically destructive.

To achieve safe ASI we need to overcome the three hurdles discussed above - correct aims, correct goals and correct implementation.

We focus predominantly on the first problem, achieving AI development which at least *aims* to be as beneficial as possible. We need to establish an environment such that those developing AGI and ASI are doing so in the interests of all people, and are provided incentives to care about safety.

For the second problem, we cannot guarantee it will be overcome via 'solving' moral philosophy. Instead looking at how one can manage moral uncertainty and disagreement is key, to insure against an AI being programmed with a mistaken conception of what the right thing to do is. It may be constructive to develop pathways whereby the interests of many people across different cultures and backgrounds can be aggregated in a democratic way that ensures the values chosen benefit everyone.

The third hurdle, correct instantiation, is a technical problem that requires collaboration between the fields of computer science, mathematics, and philosophy. The policy goals are relatively straightforward - we want to achieve ‘differential technological development’ such that safety techniques are developed before AGI. In practice this means increasing funding for safety research, while limiting the development of unsafe and/or weaponised AI.

The policies we have proposed below are largely developed from existing literature on artificial intelligence safety, particularly from the Future of Humanity Institute’s 2014 policy paper on ‘Unprecedented Technological Risks’[99]. Our suggestions are offered as examples of potential policies, illustrating the concrete actions that can be taken at this stage. We welcome suggestions, additions and revisions from policy and domain experts.

[99] Beckstead et al (2014) *Unprecedented technological risks*

5.4.2 Learning more

- Policy-makers should listen to and engage with researchers and research organisations working in the area of long-term artificial intelligence safety, including:

  - Centre for the Study of Existential Risk (CSER)
  - Future of Humanity Institute (FHI)
  - Leverhulme Centre for the Future of Intelligence (new)
  - Global Priorities Project (GPP)
  - Global Catastrophic Risks Institute (GCRI)

  It is important to establish a dialogue where policymakers listen to the suggestions of researchers, and researchers also listen to policymakers and get feedback about the political feasibility of different suggestions.

- * Horizon-scanning projects and risk registers with relevant timelines should include risks from ASI. [100] (AI experts assign 10% probability to human-level AI even by 2022)

- Commission reports/reviews on the long term risks and benefits of AI through standard government structures, for example:
  - * An independent review of the risks and benefits of ASI, e.g. on the model of the Stern Review on the Economics of Climate Change
  - * House of Commons/House of Lords Science and Technology Committee inquiry into risks from ASI
  - Government Office for Science Foresight report

- * Establish forecasting programs similar to or in collaboration with the US IARPA ACE program for predicting future global events. Include forecasting of long-term technological risks such as ASI. Techniques for knowledge creation, aggregation and elicitation like subsidised prediction markets produce expert-level judgements that update swiftly, giving more accurate information to act on.

- Create an international Intergovernmental Panel on AI risk, on the model of the IPCC, that would synthesize the evidence and produce reports for policymakers. Developing an impartially assessed knowledge base and a consensus on the importance of the issue should help create the impetus for international policy designed to promote safety and reduce possible future arms-race dynamics.

[100] Department for Business, Innovation and Skills (2014) *The allocation of science and research funding 2015/16*

5.4.3 Establish a culture of safety, collaboration and receptiveness to information

- Hold a regular conference that brings together representatives in technical, strategic, policy and legal fields, from academia, industry and government. Develop and revise milestones that would indicate progress towards ASI. Assess the current state of the field and consider what possible policy, legal and industry decisions need to be made and produce recommendations to ensure ASI is robust and beneficial. This could also be useful in the near term for spotting potential risks and actions that need taking concerning narrow AI applications, as discussed in previous sections of the paper.

- Require all companies working on AGI to have an ethics board, as Google Deepmind and Lucid AI already do.[101]

- Develop, with industry and academics, a protocol to be followed if unexpected progress is made towards self-improving AI. Require anyone working on AGI, self-improving AI or AI capable of designing new AI systems to abide by this protocol. Update this regularly. In the near term, the required responses to unexpected developments are likely to be minor, e.g. updating estimates of development timelines and re-evaluating safety priorities based on the progress made.

- Foster collaboration on safety-related research question, similar to the model of pre-competitive collaboration in the pharmaceutical industry. This involves the creation of institutions and platforms which allow companies to share data and research progress on issues that are pertinent to safety but provide little competitive advantage.

Case study: pre-competitive collaboration

An example from the pharmaceutical industry is the Accelerating Medicines Project, a venture between the US National Institute of Health, 10 biopharmaceutical companies and several non-profit organizations. Its goal is to speed up the drug development pathway by identifying efficacy and safety issues with compound collections used as starting points for many different drugs. The funding is half from industry and half from NIH, and the results are shared with the broader biomedical community.[102]

- * When creating or updating governance structures, include explicit pathways for accountability to the rights and needs of future generations. This would help to mitigate against unduly short term focus in decision-making. [from UTR paper]

[101] Notice, P. (2014) *Inside Google's mysterious ethics board*

[102] Gastfriend, E. and Lee, B. (2015) 'Pre-competitive collaboration in Pharma an overview study'

5.4.4 Differential development

- Fund technical research designed to address key problems in AI safety, as outlined in the Future of Life Institute's 'Research Priorities' document.[103] This could be through grants to institutions or individual researchers, or prizes for progress on specific problems. Much of the research needed also has useful applications to present-day technologies. It mainly falls into four domains:

  ○ Technical work in computer science, mathematics and philosophy/logic: how can we design a system so we can reliably predict its behaviour, prove its validity and security, and build in meaningful human control?
  ○ Ethics: who should have a stake in development? What goals should it be given?
  ○ Strategy and forecasting: when will different technologies be developed, and by whom? How will different actions influence development of the technology?
  ○ Policy: what policy options are available and politically feasible to maximise the benefits and minimise the risks from ASI?

- An arms race is probably the most risky way for ASI to be developed. Call for an international ban on the development and use of AI weapons, as advised in the Future of Life institute's open letter signed by over 3000 AI and robotics researchers[104].

5.4.5 Speculative suggestions for the longer term

- * Subsidise the development of safe virtual environments for AI development and testing, so that new intelligences are by default tested within such an environment.

- * Set up an anonymous contact point or hotline to enable researchers to report any concerns about ethics, safety protocols or unexpected developments in confidence.

- Fund researchers or a monitoring body to keep track of the current state of progress of different major AI projects.

5.4.6 Changes we would like to see, but do not have specific policy suggestions for

- Steer away from scenarios that would encourage extremely rapid development of AI. OpenAI (an open-source AI development project) might be an example of this, as it removes the commercial incentive structure to develop profitable technologies as fast as possible.[105] However, it creates new problems for the long run: open access to the latest progress increases the numbers of actors who could unexpectedly make progress towards ASI, and it would make things easier for groups who wanted to use AI maliciously.

[103] Aguirre et al *Research priorities for robust and beneficial artificial intelligence*

[104] *Autonomous weapons: An open letter from AI & robotics researchers* (2015)

[105] *OpenAI* (2015)

6 CONCLUSION

6.1 Summary of Recommendations

6.1.1 AI in Financial Services

AI brings a mix of evolutionary and revolutionary changes. In the near term there are a wide range of beneficial evolutions prompted by AI: a greater array of more widely available financial services, tools to improve financial decision making, and a more efficient market. These benefits are already being chased by major industry players, and it is unclear what further policy could foster these further.

The most pertinent immediate risk for advancing AI is the anticipated further development of high frequency trading. As it is unclear whether high frequency trading is a benefit or a cost, it is unclear whether more of it is a good or a bad thing. However, this uncertainty hinders effective policy intervention, and it unclear whether the best levers for policy here would be particular to AI, or wider financial services reform.

In the medium term, AI promises large disruptions of unclear sign: progressive automation (with the spectre of structural unemployment), and increasing wealth (with the spectre of increasing income inequality). However, although these changes will manifest in the financial sector, it is unclear whether they are principally issues within the financial sector.

6.1.2 AI in Healthcare

Results from early applications and the clear demand in the market demonstrate that AI has huge potential to transform medicine and healthcare in the future. There are steps that ought to be taken to stimulate the development, integration and adoption of AI technologies.

Policy 1. The ultimate aim is to have technologies that are proven to be safe and improve patient outcomes. For this to happen, it is critical that a framework of standards is developed, either within existing bodies, or set up as independent organisations that are fed into by representatives of all stakeholders.

Policy 2. Given the shifting responsibilities that would result from the increased capabilities offered by technologies including AI, medical doctors will need to be trained to understand and work with these new systems, and medical education focus will need to shift to the areas where human expertise will be key.

Policy 3. If medical AI is to achieve its fullest potential, it will need to work with other similar and complementary systems around the world. We are living in an increasingly global society, and the future of medical practice needs to not only embrace but be a key force in stimulating comparable development around the world. Global equality in adoption of medical AI will make the fullest use of the technologies available.

6.1.3 Autonomous Vehicles

Autonomous vehicle technology presents immense benefits to road safety, reduces emissions and compliments the electrification of automotive transport, and offers a life-changing option for elderly or disabled people who are currently unable to drive. As with many disruptive technologies, it presents risks as well, and we ought to take steps to mitigate such risks.

Policy 1. Facilitate relationships between the UK government, the UK's world-renowned automotive manufacturing and research industries, and university research institutions, to foster innovation in AV technology. This could involve providing funding and resources for university research labs and businesses to collaborate on, develop and test AV technologies

Policy 2. Produce a Green Paper that includes evidence for, and feedback on, possible legal obligations regarding AV cyber-security, including a legal framework for how and why certain parties or individuals are held responsible in the case of an accident. Obligations should be easily understood, are achievable in the short term and are sufficient to protect road-user safety without hindering innovation

Policy 3. Increase funding towards 'test-bed' schemes currently running in four UK cities and expand these schemes to include more cities and to serve more people; continue to publicise and explain their relevance to local residents and visitors

6.1.4 Greater-than-human intelligence

We have established that current AI research is likely to be highly beneficial and that general intelligence is still many years away. However, it is very plausible that AI greatly exceeding human capacity could be developed in the next few decades. The case is strong for acting *now* to ensure ASI is robustly beneficial. Achieving this requires both technical research and international cooperation that must be developed and successfully implemented well in advance of the development of ASI. The policies we propose to ensure ASI is beneficial are:

Policy 1. Increase information flow between industry, academics, and policymakers through conferences, government reports and intergovernmental panels where existing knowledge can be synthesized and recommendations for action can be produced

Policy 2. Use this knowledge to develop anticipatory frameworks that government, researchers and industry should abide by if, for example, unexpected progress occurs towards self-improving AI

Policy 3. Incentivise safety research by differential funding and by creating platforms to enable pre-competitive industry collaboration on safety-related issues

Policy 4. Call for an international ban on the development and use of weaponised AI

6.2 Further Discussion

6.2.1 Develop a range of anticipatory protocols

Future progress in AI is difficult to predict, but we can extract a number of general trends. Across all industries that we discuss intelligent systems have become increasingly relied upon, and there is every reason to believe this will continue. We rely on AI in a wide range of domains but only recently in those where failures could lead to injurious damage, such as transportation and healthcare. AGI raises the possibility of being able to delegate virtually all tasks traditionally carried out by human beings.

Anticipatory frameworks should be developed that are flexible enough to accommodate for developments in an unusual direction, or at an unanticipated pace. These can help to brace industry, consumers, and society as a whole for the changes that AI will bring. These must be generated on the basis of predictions made by experts in the field or, wherever possible, the planned development pipeline of companies and other organisations working on AI.

One example is making preparations to quickly update training, or bring in retraining, for those working in affected industries, such as medical practitioners when clinical decision-making is aided by AI more widely. It would also be prudent to consider the legal issues around medical AI and autonomous vehicles to avoid barriers to their adoption. Anticipatory protocols will become especially important when it comes to AGI, given we may need to act quickly by the time it is clear that it is close to being achieved.

6.2.2 Strengthen links between government, business, and academia

Improving dialogue between researchers, industry, and government should help to foster a culture of openness and safety. Not only will this help avoid undesirable outcomes of AI in areas where it poses dangers to human life, it should promote innovation and adoption across industries such that all stakeholders benefit.

If researchers and industry can efficiently communicate the details of their work in AI to policymakers, they are more likely to move funding and resources into development. This will also help policymakers to lay down anticipatory protocols for fully realising its positive effects for their constituents, as discussed above.

Greater collaboration should help avoid instances of one party antagonising another. For instance, research might be regulated invasively by government if policymakers are not properly informed in a balanced manner on developments in AI. Similarly, excessive secretiveness in industry or academia limits what governments can do to prepare for the wider consequences of disruptive technologies.

6.2.3 International coordination on AI

International coordination on AI will be valuable in the short-term when limiting its use for military purposes. By modelling our response on the regulation of biological, chemical and space-based weapons, it is possible to avoid an arms-race of autonomous weapons if universal agreement can be reached between nations. We support the Future of Life Institute's open letter on autonomous weapons from robotics and AI researchers.

As we approach human-level machine intelligence in future decades similar collaboration may be essential to avoid catastrophic outcomes. AGI created with parochial intentions could be a destructive and destabilising force. Development with international coordination will increase the likelihood that the interest of all countries, and their populace, will be looked to.

Collaboration is essential for sharing the benefits of AI in fields like healthcare between developed and developing nations. Given underdeveloped healthcare systems are often understaffed in terms of doctors, AI helpers would be especially valuable for improving health on a global level. Similarly cooperation between governments would help to brace the global market for disruptive changes in financial technologies and automation.

6.3 Concluding Remarks

Artificial Intelligence, properly instantiated and managed, holds the promise of a better future for everyone. From its applications in financial services, medicine and transport in the coming decades, to AGI and ASI in the longer term, intelligent systems may process data and manipulate the world such that it permits greater human flourishing than at any time in human history. The benefits of AI in the short-term are clear, and the pitfalls are limited to low-probability, low-impact failures. However, over longer periods of time dangers will become greater in magnitude as we defer more decision-making and actions to machines. If the reader takes away only one message from this paper, it is that we should not become stuck at either extreme when we look at the future of AI. Both reckless complacency and anxious ludditism will be destructive to human well-being, and it is evidently possible to maintain progress without sacrificing our safety.

Bibliography and Notes

1. Russell, S. J., Norvig, P. and Davis, E. (2009) *Artificial intelligence: A modern approach*. 3rd edn. United States: Prentice Hall.
2. Saletan, W. (2007) *The triumphant teamwork of humans and computers*. Available at: http://primary.slate.com/articles/health_and_science/human_nature/2007/05/chess_bump.html (Accessed: 15 January 2016)
3. Wolchover, N. (2015) *Concerns of an artificial intelligence pioneer*. Available at: https://www.quantamagazine.org/20150421-concerns-of-an-artificial-intelligence-pioneer/ (Accessed: 15 January 2016)
4. Kurzweil, R. (2006) *The singularity is near: When humans transcend biology*. Page 260. London: Gerald Duckworth & Co.
5. Bostrom, N. (1989) *How long before superintelligence?* Available at: http://www.nickbostrom.com/superintelligence.html (Accessed: 15 January 2016)
6. McCorduck, P. (1979) *Machines who think*. San Francisco: W.H.Freeman & Co.
7. *History of artificial intelligence* (2015) in *Wikipedia*. Available at: https://en.wikipedia.org/wiki/History_of_artificial_intelligence (Accessed: 15 January 2016)
8. Grace, K. (2016) *AI impacts – brain performance in FLOPS*. Available at: http://aiimpacts.org/brain-performance-in-flops/ (Accessed: 15 January 2016)
9. Chivers, T. (2011) *Japanese supercomputer 'K' is world's fastest*. Available at: http://www.telegraph.co.uk/technology/news/8586655/Japanese-supercomputer-K-is-worlds-fastest.html (Accessed: 15 January 2016)
10. *Tianhe-2 (MilkyWay-2) - TH-IVB-FEP cluster, Intel Xeon E5-2692 12C 2.200GHz, TH express-2, Intel Xeon Phi 31S1P | TOP500 supercomputer sites* (2013) Available at: http://www.top500.org/system/177999 (Accessed: 15 January 2016)
11. *Singularity is near -SIN graph - growth in supercomputer power* (2007) Available at: http://www.singularity.com/charts/page71.html (Accessed: 15 January 2016)
12. Clark, L. (2015) *DeepMind's AI is an Atari gaming pro now (wired UK)*. Available at: http://www.wired.co.uk/news/archive/2015-02/25/google-deepmind-atari (Accessed: 15 January 2016)
13. Aamoth, D. (2014) *Interview with Eugene Goostman, the fake kid who passed the Turing test*. Available at: http://time.com/2847900/eugene-goostman-turing-test/ (Accessed: 15 January 2016)
14. Thomsen, M. (2015) *Microsoft's deep learning project outperforms humans in image recognition*. Available at: http://www.forbes.com/sites/michaelthomsen/2015/02/19/microsofts-deep-learning-project-outperforms-humans-in-image-recognition/#2715e4857a0bb2ce4a431285 (Accessed: 15 January 2016)

15. BBC (2014) *Artificial music: The computers that create melodies*. Available at: http://www.bbc.com/future/story/20140808-music-like-never-heard-before (Accessed: 15 January 2016)

16. *CTO corner: Artificial intelligence use in financial services - financial services roundtable* (2015) Available at: http://fsroundtable.org/cto-corner-artificial-intelligence-use-in-financial-services/ (Accessed: 15 January 2016)

17. Schindhelm, J., TCmake_photo, koch, jonah, Feuilly, J., Brown, R. H., Roehrig, P., Malhotra, V., Solutions, C. T., Zoonar RF and AARON FEDOR 917-674-8777 (2015) *The robot and I: How new digital technologies are making smart people and businesses smarter by automating rote work*. Available at: http://www.cognizant.com/InsightsWhitepapers/the-robot-and-I-how-new-digital-technologies-are-making-smart-people-and-businesses-smarter-codex1193.pdf (Accessed: 15 January 2016)

18. Bonissone, P. P., Cheetham, W. E., Messmer, R. P. and Aggour, K. S. (2008) *Automating the underwriting of insurance applications*. Available at: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1891/1789 (Accessed: 15 January 2016)

19. Vögeli, J. (2014) *UBS turns to artificial intelligence to advise clients*. Available at: http://www.bloomberg.com/news/articles/2014-12-07/ubs-turns-to-artificial-intelligence-to-advise-wealthy-clients (Accessed: 15 January 2016)

20. Fleury, M. (2015) *How artificial intelligence is transforming the financial industry*. Available at: http://www.bbc.co.uk/news/business-34264380 (Accessed: 15 January 2016)

21. *How artificial intelligence can help banks beat back tech firms. Available at:* http://www.americanbanker.com/bankthink/how-artificial-intelligence-can-help-banks-beat-back-tech-firms-1074299-1.html (Accessed: 15 January 2016)

22. Kaushik, P., Contributors, I. and Space (2016) *Is artificial intelligence the way forward for personal finance?*. Available at: http://www.wired.com/insights/2014/02/artificial-intelligence-way-forward-personal-finance/ (Accessed: 15 January 2016)

23. Glantz, M. and Kissell, R. (2013) *Multi-asset risk modeling: Techniques for a global economy in an electronic*

24. Kensho and PR Newswire (2016) *Tony Pasquariello*. Available at: http://www.prnewswire.com/news-releases/goldman-sachs-leads-15-million-investment-in-kensho-300000102.html (Accessed: 15 January 2016)

25. *Financial services warms up to AI* (2015) Available at: http://marketsmedia.com/financial-services-warms-up-to-a (Accessed: 15 January 2016)

26. *Artificial intelligence startups see 302% funding jump in 2014* (2015) Available at: https://www.cbinsights.com/blog/artificial-intelligence-venture-capital-2014/ (Accessed: 15 January 2016)

27. *Robots and AI invade banking* (2015) Available at: http://thefinancialbrand.com/52735/robots-artificial-intelligence-ai-banking/ (Accessed: 15 January 2016)

28. *Compliance taps AI* (2015) Available at: http://marketsmedia.com/compliance-taps-ai/ (Accessed: 15 January 2016)

29. *CTO corner: Artificial intelligence use in financial services - financial services roundtable* (2015) Available at: http://fsroundtable.org/cto-corner-artificial-intelligence-use-in-financial-services/ (Accessed: 15 January 2016)

30. Hannah, Hammond, K., Andresen, L., Delgado, R., Weston, M., Gershenson, J., Krishnakumar, A. and McNulty, E. (2014) *3 reasons why banks can't afford to ignore AI*. Available at: http://dataconomy.com/3-reasons-why-banks-cant-afford-to-ignore-ai/ (Accessed: 15 January 2016)

31. *Wealth managers assess AI* (2015) Available at: http://marketsmedia.com/wealth-managers-assess-ai/ (Accessed: 15 January 2016)

32. Moore, H. and Roberts, D. (2014) *AP Twitter hack causes panic on wall street and sends Dow plunging*. Available at: http://www.theguardian.com/business/2013/apr/23/ap-tweet-hack-wall-street-freefall (Accessed: 15 January 2016)

33. Ngan, M., Images, G., Imbert, F. and foimbert (2015) *US officials: No single cause for 2014 bond market 'flash crash'*. Available at: http://www.cnbc.com/2015/07/13/us-officials-no-single-cause-for-2014-flash-crash.html (Accessed: 15 January 2016)

34. Fleury, M. (2015) *How artificial intelligence is transforming the financial industry*. Available at: http://www.bbc.co.uk/news/business-34264380 (Accessed: 15 January 2016)

35. Wyatt, E. and Bowley, G. (2014) *S.E.C. Rules would limit trading in volatile market*. Available at: http://www.nytimes.com/2010/05/19/business/19crash.html?hp (Accessed: 15 January 2016)

36. *Singapore exchange regulators change rules following crash* (no date) Available at: http://www.singaporenews.net/index.php/sid/224382417 (Accessed: 15 January 2016)

37. Berner, E. S. (2009) *Clinical decision support systems: State of the art*. Available at: https://healthit.ahrq.gov/sites/default/files/docs/page/09-0069-EF_1.pdf (Accessed: 15 January 2016)

38. Sabertehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, et al. 25-Year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the national practitioner data bank. BMJ Qual Saf. 2013;22(8):672–680

39. Charara, S. (2015) *How machine learning will take wearable data to the next level*. Available at: http://www.wareable.com/wearable-tech/machine-learning-wearable-data-sensors-2015 (Accessed: 15 January 2016)

40. Hempel, J. and media, social (2016) *Unicorns and other things we must stop talking about in 2016*. Available at: http://www.wired.com/2016/01/unicorns-and-other-things-we-must-stop-talking-about-in-2016/ (Accessed: 15 January 2016

41. Charara, S. (2015) Available at: http://www.wareable.com/headphones/the-future-of-hearables-from-fitness-trackers-to-always-on-assistant (Accessed: 15 January 2016)

42. Harris, D. (2014) *How Lumiata wants to scale medicine with machine learning and APIs*. Available at: https://gigaom.com/2014/01/08/lumiata-does-graph-analysis-for-medical-data-raises-4m-from-khosla/ (Accessed: 15 January 2016)
43. Health, I. (2015) *IMS institute on the App store*. Available at: https://itunes.apple.com/us/app/ims-institute/id625347542 (Accessed: 15 January 2016)
44. Hood, W. (2015) *A report on how doctors engage with digital technology in the workplace*. Available at: http://www.cellohealthinsight.com/wp-content/uploads/2015/11/Digital_Health_Debate_2015.pdf (Accessed: 15 January 2016)
45. Berthene (2016) *Northwestern mutual drops from the top*. Available at: https://www.mobilestrategies360.com/2015/07/23/getting-doctors-opinion-health-related-wearables (Accessed: 15 January 2016)
46. Rosenblum, A. (2015) *Your doctor may not want to see your Fitness-Tracker data | MIT technology review*. Available at: http://www.technologyreview.com/news/543716/your-doctor-doesnt-want-to-hear-about-your-fitness-tracker-data/ (Accessed: 15 January 2016)
47. *500m people will be using healthcare mobile applications in 2015* (2010) Available at: http://www.research2guidance.com/500m-people-will-be-using-healthcare-mobile-applications-in-2015/ (Accessed: 15 January 2016)
48. Lee, K. (2015) *Wearable health technology and HIPAA: What is and isn't covered*. Available at: http://searchhealthit.techtarget.com/feature/Wearable-health-technology-and-HIPAA-What-is-and-isnt-covered (Accessed: 15 January 2016)
49. *Fact sheet 39: Mobile health and fitness Apps: What are the privacy risks?* (2013) Available at: https://www.privacyrights.org/mobile-health-and-fitness-apps-what-are-privacy-risks (Accessed: 15 January 2016)
50. Timmins, N., COI and NHS (2014) *Five Year Forward View*. Available at: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf (Accessed: 15 January 2016)
51. See Appendix for clarification of the various levels of automation (as determined by NHTSA)
52. Kessler, A. M. and Vlasic, B. (2015) *Semiautonomous driving arrives, feature by feature*. Available at: http://www.nytimes.com/2015/04/03/automobiles/semiautonomous-driving-arrives-feature-by-feature.html (Accessed: 8 December 2015).
53. *Premium electric vehicles* (2015) Available at: https://www.teslamotors.com (Accessed: 8 December 2015).
54. Curtis, S. (2015) *British cities to become testbeds for driverless cars*. Available at: http://www.telegraph.co.uk/technology/news/11668491/British-cities-to-become-testbeds-for-driverless-cars.html (Accessed: 21 December 2015).
55. *Forecasts: Driverless Car Watch* (2015) Available at: http://www.driverless-future.com/?page_id=384 (Accessed: 10 December 2015).

56. Dagbladet Børsen (2015) *Elon musk – visions for Tesla, the auto industry and self-driving Teslas (interview in Denmark 2015)*. Available at: https://www.youtube.com/watch?v=bl5vLC3Xlgc (Accessed: 10 December 2015).

57. Hauser, J., GmbH, F. A. Z. and Autor (2015) *Selbstfahrende autos: Amerika schaltet auf Autopilot*. Available at: http://www.faz.net/aktuell/wirtschaft/unternehmen/verkehrsminister-foxx-selbstfahrende-autos-in-10-jahren-standard-13811022.html (Accessed: 10 December 2015).

58. *Almost one-in-10 cars 'will be driverless by 2035'* (2013) Available at: http://www.telegraph.co.uk/finance/newsbysector/transport/10545297/Almost-one-in-10-cars-will-be-driverless-by-2035.html (Accessed: 10 December 2015).

59. Lab, M. and Goddin, P. (2015) *Uber's plan for self-driving cars bigger than its taxi disruption*. Available at: http://www.govtech.com/fs/perspectives/Ubers-Plan-for-Self-Driving-Cars-Bigger-Than-Its-Taxi-Disruption.html (Accessed: 11 December 2015).

60. Neckermann, L. (2015) The Mobility Revolution: Zero Emissions, Zero Accidents, Zero Ownership. UK,: Matador.

61. The changing nature of mobility (2014) Deloitte: Global Automotive Consumer Study

62. *The Transportation Sector* (2013) Available at: http://www.c2es.org/search/common?text=transportation (Accessed: 11 December 2015)

63. This paper will not argue the case for anthropogenic climate change: such cases can easily be found online from respected institutes such as the World Health Organization, the Pew Centre on Global Warming, the International Panel on Climate Change (IPCC), and others.

64. *Driving the future today A strategy for ultra low emission vehicles in the UK* (2013) Available at: http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/239317/ultra-low-emission-vehicle-strategy.pdf (Accessed: 11 December 2015).

65. Grover, S. (2015) *10 cities aiming for 100 percent clean energy*. Available at: http://www.mnn.com/earth-matters/energy/stories/10-cities-aiming-for-100-percent-clean-energy (Accessed: 13 December 2015).

66. Hanson, M. (2012) *Electric car range anxiety*. Available at: http://www.hawaiibusiness.com/electric-car-range-anxiety/ (Accessed: 13 December 2015).

67. *Road crash statistics* (2015) Available at: http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics (Accessed: 13 December 2015).

68. Department for Transport (2014) *Annual road fatalities*. Available at: https://www.gov.uk/government/publications/annual-road-fatalities (Accessed: 13 December 2015).

69. Transport and House, W. (2014) *Casualties on London's roads at lowest level ever - transport for London*. Available at: https://tfl.gov.uk/info-for/media/press-releases/2014/june/casualties-on-london-s-roads-at-lowest-level-ever (Accessed: 13 December 2015).

70. Of the fourteen non-fatal accidents Google cars have been involved in, it has been determined independently that in all cases, a human was to blame

71. Cava, M. della (2015) *Google driverless cars in accidents again, humans at fault — again*. Available at: http://www.usatoday.com/story/tech/2015/07/01/google-self-driving-car-june-report-two-accidents-humans-at-fault-again/29592237/ (Accessed: 13 December 2015).

72. Tech Times (2015) *The Driverless car debate: How safe are autonomous vehicles?*. Available at: http://www.techtimes.com/articles/67253/20150728/driverless-cars-safe.htm (Accessed: 13 December 2015)

73. Mearian, L. (2013) *Self-driving cars could save more than 21, 700 lives, $450B a year*. Available at: http://www.computerworld.com/article/2486635/emerging-technology/self-driving-cars-could-save-more-than-21-700-lives-450b-a-year.html (Accessed: 13 December 2015)

74. Group, T. B. C., Consul, T. B. and The Boston Consulting Group (2015) *Min-sun moon*. Available at: http://www.slideshare.net/TheBostonConsultingGroup/the-road-to-autonomous-driving (Accessed: 13 December 2015)

75. Werbach, A. (2013) *The American commuter spends 38 hours a year stuck in traffic*. Available at: http://www.theatlantic.com/business/archive/2013/02/the-american-commuter-spends-38-hours-a-year-stuck-in-traffic/272905/ (Accessed: 13 December 2015)

76. Boyle, D. (2015) Black cab drivers bring central London to a standstill in Uber protest. Available at: http://www.dailymail.co.uk/news/article-3255282/Black-cab-drivers-bring-central-London-standstill-protest-TfL-licensing-hundreds-new-minicabs-week-calls-grow-crackdown-apps-like-Uber.html (Accessed: 22 December 2015).

77. Kanter, Z. and sorts, all (2015) How Uber's autonomous cars will destroy 10 Million jobs and reshape the economy by 2025. Available at: http://zackkanter.com/2015/01/23/how-ubers-autonomous-cars-will-destroy-10-million-jobs-by-2025/ (Accessed: 21 December 2015).

78. Yeomans, Lloyd's Exposure Management (2014) 'Autonomous vehicles. Handing over control: Opportunities and risks for insurance'

79. LIDAR stands for 'Light Detection and Ranging' and uses laser pulses to detect the distances of objects; it is the dominant technlogy used to map surroundings for autonomous vehicles

80. Curtis, S. (2015) *Self-driving cars can be hacked using a laser pointer*. Available at: http://www.telegraph.co.uk/technology/news/11850373/Self-driving-cars-can-be-hacked-using-a-laser-pointer.html (Accessed: 21 December 2015)

81. Ring, T. (2015) 'Connected cars – the next target for hackers', *Network Security*, 2015(11), pp. 11–16. doi: 10.1016/s1353-4858(15)30100-8

82. Zetter, K. and Espionage, C. (2015) *Researchers hacked a model S, but Tesla's already released a patch*. Available at: http://www.wired.com/2015/08/researchers-hacked-model-s-teslas-already/ (Accessed: 26 December 2015)

83. Harris, M. (2015) *FBI warns driverless cars could be used as 'lethal weapons'.* Available at: http://www.theguardian.com/technology/2014/jul/16/google-fbi-driverless-cars-leathal-weapons-autonomous (Accessed: 21 December 2015)

84. *A Roadmap for a world without drivers* (2015) Available at: https://medium.com/@alexrubalcava/a-roadmap-for-a-world-without-drivers-573aede0c968#.97xnqds8x (Accessed: 3 January 2016)

85. Etzkowitz, H. and Ranga, M. (1995) 'Triple Helix Systems:An Analytical Framework for Innovation Policy and Practice in the Knowledge Society',*Human Sciences and Technology Advanced Research Institute (H-STAR)*

86. U.S. Department of transportation releases policy on automated vehicle development | national highway traffic safety administration (NHTSA)   Available at: http://www.nhtsa.gov/About+NHTSA/Press+Releases/ci.U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development.print (Accessed: 22 December 2015).

87. Wolchover, N. (2015) *Concerns of an artificial intelligence pioneer*. Available at: https://www.quantamagazine.org/20150421-concerns-of-an-artificial-intelligence-pioneer/ (Accessed: 15 January 2016)

88. Mills, J. (2015) *'Robots we design could crush humanity like an anthill', Stephen hawking warns*. Available at: http://metro.co.uk/2015/10/08/artificial-intelligence-we-design-could-crush-humanity-as-easily-as-an-anthill-stephen-hawking-warns-5429717/ (Accessed: 15 January 2016)

89. Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Figure 8, Page 70. United Kingdom: Oxford University Press.

90. CRASSH Cambridge (2015) *Professor Stuart Russell - the long-term future of (artificial) intelligence*. Available at: https://www.youtube.com/watch?v=GYQrNfSmQ0M (Accessed: 15 January 2016)

91. Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Chapter 4. United Kingdom: Oxford University Press.

92. Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Page 63, Figure 7. United Kingdom: Oxford University Press.

93. Bostrom, N. (2013) *Superintelligence: Paths, dangers, strategies*. Page 247, Box 13 'A risk-race to the bottom'. United Kingdom: Oxford University Press.

94. Shalal, A. (2016) *Pentagon eyes $12-15 billion for early work on new technologies*. Available at: http://www.reuters.com/article/us-usa-military-technologies-idUSKBN0TX1UR20151214 (Accessed: 15 January 2016)

95. Williams, E. G. (2015) 'The possibility of an ongoing moral catastrophe', *Ethical Theory and Moral Practice*, 18(5), pp. 971–982. doi: 10.1007/s10677-015-9567-7

96. Soares, N. (no date) *The value learning problem*. Available at: https://intelligence.org/files/ValueLearningProblem.pdf (Accessed: 15 January 2016)

97. Bird, J. and Layzell, P. (no date) *The evolved radio and its implications for Modelling the evolution of novel sensors*. Available at: https://people.duke.edu/~ng46/topics/evolved-radio.pdf (Accessed: 15 January 2016)

98. *Nuclear weapon* (2016) in *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Nuclear_weapon (Accessed: 15 January 2016)

99. Beckstead, N., Bostrom, N., Bowerman, N., Cotton-Barratt, O., MacAskill, W., O hEigeartaigh, S. and Ord, T. (2014) *Unprecedented technological risks*. Available at: http://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf (Accessed: 15 January 2016)

100. Department for Business, Innovation and Skills (2014) *The allocation of science and research funding 2015/16 Investing in world-class science and research*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/332767/bis-14-750-science-research-funding-allocations-2015-2016-corrected.pdf (Accessed: 15 January 2016)

101. Notice, P. (2014) *Inside Google's mysterious ethics board*. Available at: http://www.forbes.com/sites/privacynotice/2014/02/03/inside-googles-mysterious-ethics-board/2/ (Accessed: 15 January 2016)

102. Gastfriend, E. and Lee, B. (2015) P*re-competitive collaboration in Pharma an overview study*. Available at http://futureoflife.org/data/documents/PreCompetitiveCollaborationInPharmaIndustry.pdf (Accessed: 18 April 2016)

103. Aguirre, A., Brynjolfsson, E., Calo, R., Dietterich, T., George, D., Hibbard, B., Hassabis, D., Horvitz, E., Kaelbling, L. P., Manyika, J., Muehlhauser, L., Osborne, M., Parkes, D., Roff, H., Rossi, F., Selman, B. and Shanahan, M. (no date) *Research priorities for robust and beneficial artificial intelligence*. Available at: http://futureoflife.org/data/documents/research_priorities.pdf (Accessed: 15 January 2016)

104. *Autonomous weapons: An open letter from AI & robotics researchers* (2015) Available at: http://futureoflife.org/open-letter-autonomous-weapons/ (Accessed: 15 January 2016)

105. *OpenAI* (2015) Available at: https://openai.com/blog/introducing-openai/ (Accessed: 15 January 2016)