The 'AI problem': A Public Policy perspective

Author: Jean-Paul Skeete

Keywords: Artificial Intelligence; Public Policy; Black Box; AGI; Explainability; Regulation

Abstract

As novel applications of Artificial Intelligence (AI) increasingly populate our newsfeeds and social media timelines, the near-future implications of this technology have increasingly become the topic of casual conversation among friends and family. In fact, one could argue that the hype and cultural relevance around AI in 2019 is at an all-time high. It is therefore imperative that we in the academy of social sciences, continue to pursue grounded discussions in all areas relevant to the application of artificial intelligence, and the implications that will likely follow.

In-depth explorations of our strategic situation will subsequently help us identify more effective interventions. Strategic analysis is especially crucial when we are radically uncertain about not only the fundamental qualities of AI, but also which directions of change would be desirable as opposed to detrimental. This paper's principal aim, therefore, is to join the interdisciplinary search for crucial considerations: arguments that can potentially change our views at any scale, from the fine details of implementation to the general topology of desirability.

There are no established methods for conducting research in this new frontier, and thus difficult original thinking will be necessary. The sooner we can – at the non-specialist level – understand the relevant basics around AI, the better equipped we will be to address the implications for society's future wellbeing. This paper aims therefore to present various AI concepts and concerns, that they may be interrogated among peers while we still have time, as we begin to contend with the accelerating emergence of this powerful, underlying technology.

1. Introduction

It has been argued that in today's society, the nature of any function that embeds logic is rapidly transitioning from 'atoms to bits', and the root cause of this transition is the ubiquity of end-user microprocessors found in a wide variety of products such as cars, mobile phones and insulin pumps (Ohm & Reid, 2016). The transition from hardware to software has given rise to devices which have become collectively known as the 'internet of things' (IoT). However, as with most cutting-edge technology, regulators in their respective sectors are beginning to face new and unique challenges that were previously not manifest.

This paper aims to highlight some of the most pressing concerns in various industries around the regulation of software, in order to better understand the implications that more advanced Artificial Intelligence (AI) applications will present in the not-so-distant future. Section 2 lays out some interdisciplinary theoretical underpinnings, Section 3 highlights several interesting industry applications of AI in the form of abbreviated 'case reviews', and section 4 conducts some critical analysis on a variety of AI issues currently being debated. Section 5 concludes with possible interventions and closing comments. This paper argues that as the global trend of industrial digitization accelerates, scholars will need to increasingly devote resources to examining the many implications – known and unknown – of this most significant sociotechnical transition.

2. Theoretical underpinnings

2.1. Regulating 1's and 0's

Regulating increasingly data-driven technologies is no easy task, even when using direct interventions in the form of Command and Control (CAC) regulations, as seen in the Volkswagen "dieselgate" scandal (Skeete, 2017). Nevertheless, as regulators begin to contend with limited applications of AI, e.g. autonomous vehicles in the automotive industry (Taeihagh & Lim, 2019), they are faced with emerging issues of liability, data access and type approval (Skeete, 2018). However, the common perception remains that

when it comes to cutting edge technologies, regulators are mostly playing 'catch up' in these unregulated spaces. As one blogger so eloquently put it "technology changes so fast that new regulations tend to protect yesterday from last Thursday" (Searls, 2019).

Fundamental regulatory logic dictates that a system's operation can be observed and verified, however the increasing presence of software in regulated artefacts disrupts this logic in several ways. For starters, software is easier to modify than hardware. While hardware requires cumbersome and visual alterations, software only needs the addition of a few lines of code to alter its behaviour. Jonathan Zittrain (2005) described software as a 'generative technology' that allows unmodified hardware to execute new algorithms (functions) without the need for 'bolting in' new parts. While software's adaptable and accessible nature makes it easy to master by many, it can also be easily obscured in binary code, tamper-proof chips or via encryption. Thus, software's generativity and obscurity pose a problem for regulators as they transition from regulating products to regulating code.

2.2. What is Artificial Intelligence?

For simplicity's sake, intelligence is defined in this paper as the ability to accomplish complex goals (Tegmark, 2017). In humans, this manifests as biological intelligence, but when it is engineered (non-biological), it is referred to as Artificial Intelligence or 'AI' and a subset of this research discipline is Machine Learning (ML), which enables computers to learn without being explicitly programmed. This paper is mostly interested in the subset of ML that is responsible for today's explosion in AI development: Deep Learning. This is a computer's ability to 'learn' underlying features in data using Artificial Neural Networks or 'neural nets' more simply.

Narrow AI is a system that can accomplish a narrow set of goals, e.g., play chess or drive a car while Strong AI is system that can accomplish any cognitive task as well as humans and is commonly referred to as Artificial General Intelligence (AGI). Many believe that AGI will inevitably lead to "Superintelligence", which is a form of general intelligence well beyond that of any human. While AGI does not yet exist, it is this version of AI that causes the most concern to those who believe that AI has the potential to pose a 'non-trivial' threat to humanity (Bostrom, 2014; Guihot, Matthew, & Suzor, 2017). While this paper's

focus is on the short and medium-term implications of AI deployment, it must be noted that many dismiss the possibility of AGI as a distraction and fanciful futurism. Let us not forget however, that Nobel laureate Ernest Rutherford, known as the "father of nuclear physics", said in 1933 that nuclear energy was "moonshine". This was less than twenty-four hours before Leo Szilard invented of the nuclear chain reaction. AI safety research is not futurism, it is simply the recognition that the risks of AI are non-negligible (Bostrom, 2014).

2.2. Deep Learning and Artificial Neural Networks

Deep Learning in its current state of development is a metaphorical and literal "black box" that takes one or multiple inputs, such as the data from the sensors in driverless vehicles (CAVs[1]), and processes those inputs into outputs, which are the controls of the vehicle. In other words, Deep Learning uses neural nets to perform a function that takes in data inputs, and 'spits out solutions' as data outputs.

2.3. The Deep Learning 'Black Box'

As deep learning is deployed more widely, neural nets have already begun solving many of society's problems such as facial recognition along the EU border, language translation, spotting deadly tumours, making multi-million-dollar financial trades on the stock market, and the list goes on. The reality is however, that 'no one really knows how the most advanced algorithms do what they do [and] that could be a problem' (Knight, 2017).

AI precision is bundled with opaqueness and a certain interpretive cost that made these technologies be commonly referred to as "black-box" systems. Code is often kept undisclosed and is fundamentally difficult (even impossible) to understand. The type of data that is gathered, the associations that are targeted, and the concerns that are factored into the algorithmic predictions are not at all obvious. These layers of opacity can disguise biased, discriminatory or otherwise undesirable results from supervision until negative results become obvious (Packin, 2017).

As institutions move away from decipherable mathematical models and increasingly adopt complex machine-learning approaches, financial, medical and military automated

[1] CAV is short for Connected and Autonomous Vehicle, as it is referred to in the UK

decision-making runs the very real risk of becoming 'inscrutable'. Joel Dudley, lead researcher at a medical AI firm was quoted as saying:

> We can build these models, but we don't know how they work" (Knight, 2017).

The problem is as follows: While "black box" neural nets can approximate any function, studying the structure of the black box won't reveal any insight into the structure of the function that is being approximated. Also, from a traditional statistics viewpoint, neural nets are non-identifiable models, i.e. given a dataset and network topology, there can be two neural nets with different variables that produce the same result. This makes analysis quite difficult. Consider the following quote from a MIT (AI) Professor Tommi Jaakkola:

> If you had a very small neural network, you might be able to understand it, but once it becomes very large, and it has thousands of units per layer and maybe hundreds of layers, then it becomes quite un-understandable. (Knight, 2017).

This obscurity (and secrecy) protects businesses and other public entities against open disapproval, making it harder to identify the role of human judgment, and mandate it as part of the process in when necessary (Packin, 2017). Subsequently, issues have begun to emerge in the banking, healthcare and military sectors (Brooks, 2015), and has even given rise to the sub-field of "XAI", or 'explainable artificial intelligence' (DARPA, 2018).

3. Case Reviews: Industrial AI applications

This section presents three succinct case reviews within the banking, automotive and military sectors that make clear the real-world challenges of AI deployment and facilitates more fact-based discussions in the sections that follow.

3.1. Banking

The financial sector is one of the best industries to explore the regulation of software for a few reasons. First off, the banking sector *the* prime purchaser of IT services and products globally. For example, one third of Goldman Sachs' 33,000 staff are engineers, which is more than LinkedIn, Twitter, Facebook (Arner, Barberis, & Buckley, 2015). In fact, the banking and securities sector's IT budget as a percentage of its revenue (7%) surpasses

that of even the technology and communication industry (3.7%) (Deloitte, 2017). In the first quarter of 2017, Backrock, the world's largest asset manager ($5tn in assets) fired several prominent stock-picking fund managers in favour of pursuing more "quantitative investment strategies" (Foley, 2017) or AI. One consulting firm estimates that by 2025, AI agents[2] will replace roughly 10% (230,000) of the global human work force within capital markets, with money management professionals accounting for 40% of those losses (Opimas, 2017).

The second major reason for the financial sector's relevance to AI is the regulatory and technological transformation it underwent immediately after the 2008 Global Financial Crisis (GFC). Post-GFC, regulators imposed heavy regulatory burdens (Dodd Frank, Basel III) on banks as regulators realized they could no longer rely on the internal risk management systems within these institutions. At the same time, disruptive financial technologies were developing quickly, especially in the area of service automation such as online mortgages and wealth management 'robo-advisors' (Buchak, Matvos, Piskorski, & Seru, 2018; Lee & Shin, 2018). This perfect storm of regulations and disruptive innovation is credited with giving birth of newcomer financial technology (Fintech) firms.

3.1.1 Regtech

Regulatory technology or 'Regtech' is a subset of Fintech, developed in response to regulatory burden to automate reporting and compliance processes in the banking industry, and the increasingly data-driven nature of financial services. The proposed benefits of Regtech include high resolution risk assessment, quasi real-time monitoring of firm behaviour and market outcomes (FCA, 2019), and better macroprudential policy-making via 'early warning systems' that improve systemic financial stability. For example, Central Banks are using data 'heat maps' that highlight potential problems based on big data streams, which is exactly where AI excels; pattern recognition and probabilistic reasoning. However, as Fintech moves from the digitization of money to the monetization of data, financial regulatory frameworks (and Regtech) must evolve to consider previously unnecessary notions of data sovereignty and algorithm supervision.

[2] An agent is anything capable of altering the world, including chemical agents (Bryson & Theodorou, 2018).

Regtech may also be used in banking operations as a built-in limiter to Fintech applications, where a trading strategy for example has Regtech parameters that limit the set of available actions by excluding those that are deemed unlawful (Enriques, 2017). This has been seen in EU algorithmic trading[3], where due to the pervasiveness of 'algo trading' in securities markets, policy makers have had to include 'algo traders' within the oversight perimeter. This raises an important question: If product governance in this sector is essentially code governance, should product governance extend to coders and software developers themselves? (Enriques, 2017).

It is also important to note here that capital requirements[4] regulations are based on the idea that shareholders and depositors cannot observe a firm's risk-taking actions. Some argue that Regtech could signal the end of asymmetric information, where disparities in knowledge between buyers and sellers is eliminated. Banks could place their accounts on "trustless" legers (e.g. Blockchain) that are observable to the public, or alternatively, keep them private but monitored by publicly observable algorithms that track the performance and solvency of the firm.

3.1.2. Blockchain

'Blockchain' is a distributed ledger technology that is showing potential as a Regtech application in banking, specifically in the areas of Know Your Customer (KYC) and Anti-Money Laundering (AML) compliance. Blockchain can help in areas such as client identification and verification via digital identities, customer screening and customer risk analysis (e.g. negative press, customer profession etc.) as well as tax reporting (fiscal residence) to name a few (Lootsma, 2017).

As mentioned above, after the financial crisis, regulators no longer rely on the internal risk management of institutions and with good reason. For example, the majority of trading in major securities occurs off-exchange via Electronic Communication Networks (ECNs) or 'dark pools'. However, the US and EU are moving to require mandatory

[3] Algorithmic trading is a method of executing a large order using automated pre-programmed trading instructions. 'Algo trading' as it is known, was developed so that traders would not need to constantly monitor stocks and repeatedly send out orders manually.

[4] A capital requirement is the amount of capital a bank or other financial institution has to hold as required by its financial regulator.

reporting of all transactions in listed securities irrespective of the transaction platform (ECN, major exchanges, etc) (Arner, Barberis, & Buckley, 2016b). Regulatory arbitrage (moving activities to unregulated environments to avoid scrutiny) and excessive reliance on financial institutions' internal quantitative risk management systems have been identified as the root causes of the GFC and explains the regulatory focus on these risks today. Thus governance technologies like Regtech and Blockchain that facilitate monitoring, coordination and exchange, can potentially expand the boundaries of effective regulatory interventions in digital finance (Allen & Berg, 2018).

3.2. Automotive

In 2015, several automakers, the most prominent of which was Volkswagen (VW), deceived US and EU regulators by fitting 'defeat devices' - developed by Bosch, a Tier 1 automotive supplier - into several models of their diesel passenger vehicles. The result was that during real world driving, toxic Nitrogen Oxide (NO_x) emissions were up to 40 times higher than Volkswagen's stated test results. While policy failures in the US and EU (Skeete, 2017; Zachariadis, 2016) have been previously explored, this section focuses instead on the technical nature of the 'Dieselgate' deception.

The main culprit in this scandal is the EDC17 diesel engine control unit (ECU) designed and supplied by Bosch. It was revealed that Bosch created the defeat function in the ECU, and VW subsequently enabled it in its vehicles. In simple terms, when the vehicle was switched on, the EDC17 ECU used environmental parameters such as time elapsed, distance travelled and steering wheel angle checks to detect whether or not an emissions test was being conducted. If the ECU detected no test conditions, it would disable the vehicle's emission control measures, resulting in significantly higher CO_2 and NOx levels (Contag et al., 2017).

While emissions test conditions are standardized and public for transparency's sake, it allowed car manufacturers (OEMs) to intentionally alter the vehicle's behaviour during the test cycle, a practice known as 'cycle beating'. A premium class vehicle has more than 70 electronic units and over 100M lines of code, and nearly all aspects of the engine's operation are controlled by the ECU. It is important to note that while *test conditions* are

transparent, laboratory *test data* from individual vehicles is not available to third parties (Transport & Environment, 2019). Thus, this black box aspect of the testing makes it nearly impossible to uncover the presence of a software-based defeat device during a test, forcing regulators to rely on heavy fines (Contag et al., 2017). The incompatibility between black box testing and modern software assurance highlights a critical research agenda going forward as regulators oversee and evaluate increasingly complex systems. Concrete examples like this are key to grounding these discussions and make clear the real-world difficulties faced by regulators (Contag et al., 2017).

Another difficulty was that the Bosch defeat device was adversarial, meaning that it was not a bug or flaw in the code, but an intentional alteration of the system's behaviour under test conditions, making it more difficult to detect. Fortunately, the aftermarket 'tuner' community – car enthusiasts who regularly modify ECU settings for performance gains – were able to help researchers fully access the EDC17 ECU.

Contag notes that (2017, p. 16) "as software control becomes a pervasive feature of complex systems, regulators in the automotive domain (as well as many others) will be faced with certifying software systems whose manufacturers have an immense financial incentive to cheat."

While this case review does not feature the use of advanced AI systems, it clearly illustrates the severe consequences that can result from the deployment of a simple adversarial algorithm. According to a joint report published by the Greater London Authority and Transport of London, diesel exhaust is a major contributor of air pollution in London that prematurely kills approximately 9,500 people each year (Skeete, 2017).

3.3. Military

Over the past few years, significant debate has emerged around the development and future deployment of Lethal Autonomous Weapons Systems (LAWS) by military forces around the world. Many argue that LAWS pose an unacceptably high risk to human well-being (Guihot et al., 2017), with some even raising distant-future concerns around 'killer robots'. As with the previous example of the automotive industry, it is important that

these discussions are grounded in facts, and a good place to start is with properly defining autonomous systems, which is not the same as automated systems (M. L. Cummings, 2019).

Automated systems are guided by clear repeatable rules based on unambiguous sensed data. Autonomous systems on the other hand process raw, unstructured data about world around them to generate information, and make decisions in the face of uncertainty. This set of capabilities is often referred to as self-governing. Some systems however are neither fully automated nor fully autonomous, but somewhere in between. One example of this is the US Air Force Global Hawk, which is highly automated but can invoke low-level autonomy if its communication link to remote operators is disrupted and can land itself at an emergency airfield.

Proponents of a ban on LAWS advocate for 'meaningful human control' (MHC), however this is an ambiguous term with a wide variety of interpretations. Does MHC mean that a human must initiate the launch? A person sitting in a chair 4000 miles away from the point of weapon release hardly qualifies as MHC. Does MHC mean a human has to monitor a weapon until impact, and possibly have the ability to abort the mission? In such cases MHC tasks are highly prone to human error and can have deadly unintended consequences.

Engineers constantly struggle to determine which functions and roles should be automated/autonomous vs. those that should be assigned to humans. In fact, it is often not obvious whether a human or technology should be in control when designing certain systems such as autonomous vehicles, certain medical devices and LAWS. These design gaps usually lead to discussions about the ethics and social impacts of these technologies, especially those that are safety-critical. Managing uncertainty means understanding how much MHC can and should be applied in a given scenario, and the implications of using LAWS in such a setting (M. L. Cummings, 2019).

3.3.1. LAWS in 'high-uncertainty' engagements

The main problem in high uncertainty targeting scenarios is time pressure. Identification and classification of targets is the top priority, however when moving targets are perceived as a threat, time is limited for human controllers (e.g. fighter pilots) to make high quality

decisions. Research has shown that 'warfighters' are susceptible to several psychological biases that inhibits their ability to evaluate all relevant information (M. L. Cummings, 2019). This has led to disastrous outcomes as was seen in the case in the Chinese embassy bombing in Yugoslavia in 1999 (Rasmussen, 2007) and Operation Provide Comfort in Northern Iraq in 1994 (GAO, 1997).

Autonomous weapons in theory seem well-suited to assist human operators with decision making in time-critical targeting situations, where humans' ability to quickly process unbiased information is limited. This argument has been supported by citing automated (not autonomous) defensive weapons systems such as Patriot missile system or Israel's Iron Dome. Humans are limited in their ability to respond to incoming rocket attacks due to biological factors such as neuromuscular lag, where an attentive human takes about half a second to see a problem and respond accordingly. For this reason, automated and autonomous *defensive* systems have been entrusted with the ability to authorize weapons launch.

Currently, there are no (known) applications of autonomous *offensive* systems (LAWS) in dynamic battlefield scenarios. Automated offensive weapons, however, do exist and are routinely used, such as the US Air Force's missile carrying Predator drone. Predator and Reaper drones carry out the same tasks as combat aircraft, with humans supervising from a safe distance instead of in the cockpit.

Target identification is currently assigned to human operators; but the U.S. military has devoted significant resources over the past two decades to the development of automated target recognition systems for unmapped and dynamic targets. Progress, however, has thus far has been slow and difficult despite the integration of computer vision and machine learning, resulting in high false positive rates and other technical difficulties. Some argue that computer vision and ML are proving to be an Achilles heel for military (LAWS) and civilian (AVs) autonomous systems, which have been thus far limited in their ability to integrate incoming sensor data under dynamic or unfamiliar conditions, and subsequently make high quality decisions.

Given these issues, LAWS are currently unable to reliably 'make sense' of the world, especially in dynamic conditions, and thus any weapons system (offensive or defensive)

that uses autonomous reasoning based on ML will be deeply flawed. Furthermore, the exploitation of computer vision weaknesses raises cybersecurity concerns. This has been seen in printed eyeglass frames that allow persons to evade or deceive state-of-the-art facial recognition systems (Sharif, Bhagavatula, Bauer, & Reiter, 2016), or printed posters and stickers that are used as 'visual adversarial perturbations' to successfully defeat autonomous vehicle road sign classification systems (Eykholt et al., 2017). This is concerning given that these two examples can corrupt safety-critical AI applications.

Due to these deep technological flaws, target acquisition remains the exclusive responsibility of human operators in current military operations.

3.3.2. Meaningful Human Certification

In missions that prosecute static targets, the pilot does not pick the target (this is done by a team of people), nor do they authorize the target (this is done by a senior official in consultation with a team of lawyers). A pilot bombing a predesignated target is there to deliver and confirm the weapon's release in accordance with the rules of engagement. In these missions, most MHC comes from the point of final target approval. While combat pilots can execute some judgement and abort a mission if some parameters are not met, such as civilians in the area, does it qualify as MHC in these kinds of missions? Due to improvements in technology, pilots in these types of missions can and have been successfully replaced by smart missiles where from over 1000 miles away, the hybrid autonomous/automated Tomahawk missile can deliver similar payloads with meter-precision accuracy. The Tomahawk can also 'loiter' for hours over the battlefield, allowing commanders to redirect the missile if circumstances change. Using digitized scene mapping from computer vision (not ML), Tomahawk missile precision far surpasses that of any human pilot and has spawned the term "surgical strike". These missions are expensive however with missiles costing ~$1M per copy. There also exists "fire-and-forget" weapons that are beyond any real time MHC, that have replaced error-prone operators, where mistakes occur only when humans incorrectly enter target information or select the wrong target.

As is, MHC is ill-defined when applied to any kind of weapons system (autonomous or human-operated) as the 'fog of war', time pressures, and imperfect information all

converge to make real-time decision making on the battlefield extremely difficult and prone to error. Thus, the most meaningful forms of human control are the a priori decisions made about which targets to prosecute (building, person, etc.) and under what conditions.

Alternatively, could a more meaningful discussion emerge around the concept of Meaningful Human Certification (MHCrt)? The idea is that the use of offensive LAWS could be based on a two-step certification process. The first human strategic layer involves a high-level decision maker determining the target, and the second layer would be the deployment of an appropriate autonomous system capable of identifying and engaging said target with better-than-human odds. Unfortunately, certification of autonomous systems with better-than-human performance at complex safety-critical tasks has not been solved for either military or civilian applications and remains a point of intense debate.

Because these systems rely on probabilistic reasoning, powered by black box algorithms running through neural networks, it is hard to predict with high levels of confidence how these systems will perform in dynamic environments. Compounding the issue is that engineers have yet to devise a reliable way of testing these systems for errors of commission and omission. If LAWS is to become a reality, then there must exist strict certification criteria for strategic target selection by humans, and target identification and engagement by LAWS. MHCrt should involve rigorous, objective testing that clearly demonstrates better-than-human performance in comparable circumstances, while effectively safeguarding against cybersecurity attacks.

There exist other barriers to the responsible design and deployment of LAWS, like the fact that global military equipment manufacturers are indemnified against incidents on the battlefield. This policy is incompatible with the responsible use of LAWS in theatres of war, and manufacturers of these weapons, and the branches of the military that use them should be liable for potential misuses and abuses of these technologies. Some form of MHCrt would go a long way towards increasing accountability.

Those who argue against LAWS (e.g. Stop Killer Robots campaign, Future of Life Institute) often liken them to nuclear weapons, citing lack of proportionality and

distinction which are core tenets of Just War Theory. However, LAWS inherently demonstrate the opposite characteristics, where the Tomahawk missile for example, is the most precise and proportional weapon within the known US military arsenal. While it is possible for non-state actors to abuse these technologies, nuclear weapons and land mines have demonstrated that formalized bans do not deter rogue actors.

The argument has been made that given what we know about human error, and the lack of any real MHC in many documented weapons launch cases, it would be more ethical for commanders to launch a certified LAWS. This capability currently exists with regards to static targets (e.g. buildings), where it is preferable to launch a Tomahawk missile at a legitimate military target than deploy a human-piloted combat aircraft. It must conceded that there is a risk that political leaders would be less inclined to engage in diplomacy, which is why MHCrt focuses just as much on strategic human certification as it does on technological design certification.

While LAWS are not ready to negotiate dynamic environments today, Cummins argues that (2019) we owe it to potential victims of human error, and to pilots and operators likely to make mistakes in times of war. She encourages military researchers to continue working towards balanced role allocation in weapons systems that are at the very minimum, less prone to error than humans. The future of responsible LAWS is likely the implementation of meaningful human certification, "not insisting on an illusionary concept of meaningful human control" (M. L. Cummings, 2019).

A related concern is that in the highly competitive market for highly skilled roboticists and related engineers, aerospace and defence funding is dwarfed by the economic resources available to the more lucrative commercial sectors interested in AI (automotive, banking, IT). As a result, there is a widening gap between the global defence industry and its commercial counterparts in terms of AI technological innovation, as the best and brightest engineers are siphoned away by commercial interests. (M. 'Missy' L. Cummings, 2017)

4. The 'AI problem': A discussion

The following discussions on AI in this section aims to briefly highlight some concerns that have been voiced about AI development in general, many of which are relevant to the cases reviewed above. The topics below are ordered from those this paper considers to be the most immediate to the most distant concerns, which more or less correlates with the overall maturity of the technology.

4.1. "The death of expertise and rise of data science"

Some argue that just as the Industrial Revolution caused socioeconomic upheaval whereby mechanization reduced the need for human manual labour in manufacturing and agriculture, Al systems will reduce the demand for human labour in the service sector (Scherer, 2015). The most common retort is that much like today, there will be an abundance of new professions that will replace the old ones lost.

A sobering fact, however, is that the vast majority of occupations today already existed a century ago, and when we sort them by the number of jobs they provide, we have to go down to 21st place on this list until we encounter a new occupation - software developers - who make up less than 1% of the US job market (Tegmark, 2017).

AI systems perform tasks that once were the exclusive province of well-educated humans and is described by Hans Moravec's "landscape of human competence" (Moravec, 1998) where elevation represents difficulty for computers, and the rising sea level represents what computers are able to do. In his own words:

> Computers are universal machines, their potential extends uniformly over a boundless expanse of tasks. Human potentials, on the other hand, are strong in areas long important for survival, but weak in things far removed. Imagine a "landscape of human competence," having lowlands with labels like "arithmetic" and "rote memorization," foothills like "theorem proving" and "chess playing," and high mountain peaks labelled "locomotion," "hand-eye coordination" and "social interaction." Advancing computer performance is like water slowly flooding the landscape. A half century ago it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has

> reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century. I propose that we build Arks as that day nears, and adopt a seafaring life! - (Moravec, 1998)

4.2. AI development: Concentration of resources, 'compute power' and asymmetric knowledge

Industry trends suggest that AI development (as with most twentieth century technologies), will be largely driven by commercial entities rather than state agencies or small private actors. AI's commercial potential has already sparked a veritable AI arms race as large firms aggressively invest in various AI projects (Scherer, 2015).

Three of the four largest tech firms in the world (Galloway, 2018) are also ranked among the world's 8 largest AI companies (Forbes, 2017), with Google even appearing twice on that list[5]. Most companies on the list are approaching $1 trillion in market capitalization worth (individually), with the 'poorest' company – Baidu – valued at $99 billion. There is however one non-profit among them, OpenAI, but it was co-founded and is backed by the founders of several multi-billion-dollar companies including Tesla, PayPal, LinkedIn, and Y Combinator, as well as being supported by Amazon. UPDATE: While this paper was being written, OpenAI made a breakthrough with their GPT-2 AI text generator, refused to release the entirety of their research findings (as they usually have) citing 'responsible disclosure' for fears of misuse, and subsequently turned themselves into a for-profit enterprise (Waters, 2019).

It appears then, that the centre of gravity for AI development is to be found in the same place as the public risks of the twentieth century - large, highly visible corporations. If this turns out to be true, then the most significant advances in AI will likely come from regime-dominant, incumbent firms that are highly visible to courts and regulators. Established firms' economies of scale and access to greater financial and human capital is even more concerning given recent indications that computational or 'compute' power (Tkatch,

[5] In order Deepmind (owned by Google), Google, Facebook, OpenAI, Baidu, Microsoft, Apple, and IBM.

2019) is a crucial – and likely the biggest determining - factor in developing more sophisticated AI (Scherer, 2015). Incumbent accumulation (Berggren, Magnusson, & Sushandoyo, 2015) also exacerbates the opacity problem because private firms are motivated to maintain secrecy, and not compelled to share information. In fact firms benefit from the law of patents that protect their legitimate interests in new technologies (Guihot et al., 2017).

Another consideration is knowledge asymmetry. While regulators possess an advantage in expertise over legislators and judges who are most often generalists, with workloads that span a variety of industries and several fields of law, agency expertise pales in comparison to that of large firms developing cutting edge AI applications (Scherer, 2015). Even the notion of expertise becomes quite 'slippery' when we consider that dominant strands of AI research have changed frequently during the industry's six decades of existence. Due to the broad interdisciplinary nature of AI research, could an IT engineer with no ML training be considered an 'expert'?

Probably for the first time in history, governments find themselves at an overwhelmingly disproportionate disadvantage against the major corporate stakeholders in AI. (Guihot et al., 2017)

4.3. AI Regulation: Ex-ante intervention, uncertainty and regulatory delay

Regulators enjoy the freedom to conduct independent investigations and share legislatures' ability to act ex ante (before the event) compared to courts who by contrast, are inherently reactive institutions. Despite ex-ante action being diminished by rapid changes in direction and scope of AI research, regulators can still be effective by adopting standards laying out the characteristics that AI systems should have (Scherer, 2015), such as being limited to certain activities or remaining susceptible to meaningful human control. Regulating ex-ante can be used to avoid or limit risks to human health and safety, the environment, or mitigate against some moral hazard (e.g. gene manipulation). However, AI applications that pose a low risk in one area (e.g. human life) may simultaneously expose segments of society to other significant risks (e.g. loss of privacy, unemployment) (Guihot et al., 2017).

Many argue that regulators should avoid regulating purely on the threat of unknown future risks and favour "wait and see" regulatory style approaches instead. This perspective is very much in alignment with other concepts such as 'permissionless innovation' that allow markets more freedom develop. This outlook found favour in the banking sector pre-GFC and remains popular elsewhere, as is the case with the Pharmaceutical industry, which is a 'wait-and-see industry', regulated by clinical trials. Ultimately, we must remain cautious about regulatory inaction "because a probability of harm is, under many circumstances, a sufficient reason to act. (Sunstein, 2003, p. 1055)"

Regulators should keep every available option open, because after over sixty years of AI research and development, regulatory intervention at this time could not be criticized as being overly reactive or precautionary (Guihot et al., 2017).

4.4. Exponential rate of change

Today's three largest global money market funds (MMFs) were established between the 1940's and 1970's, but in 2014, Alibaba launched an online-only MMF, and in 9 months it became the world's 4th largest MMF. Exponential growth in digital technologies create metaphorical 'jumpgates', that allow firms to progress from 'too-small-to-care' directly to 'too-big-to-fail', completely bypassing the 'too-large-to-ignore' phase (Arner, Barberis, & Buckley, 2016a). This presents a significant challenge for regulators as developments in AI will likely outpace any attempt at regulating it, given that the 'too-large-to-ignore' window is where most regulatory intervention traditionally occurs.

Thus, the consequence of exponential rates of change is limited opportunities at significant intervention. Traditionally, regulators have relied on trial and error in order to keep technology beneficial, i.e. learning from mistakes. For example, we invented cars, repeatedly crashed, then invented seatbelts, airbags and now autonomous vehicles.

Unfortunately, more powerful the technology, the more likely a single accident will be devastating enough outweigh its benefits (e.g. nuclear war). Hence as we become more technologically sophisticated, the less we should rely on trial and error approaches to safety engineering. As AI controlled systems become more ubiquitous, the consequences of an AI malfunction crashing critical infrastructure such as the stock market, power grid,

or nuclear weapons system becomes non-trivial, and should prompt regulators to become more proactive and less reactive (Tegmark, 2017).

4.5. Black box architecture, liability and the explainability dilemma

As discussed above, machine learning technology is inherently opaque, even to computer scientists. This does not mean that all future AI techniques will be equally unknowable, but deep learning, by its very nature, is a particularly inaccessible black box. Looking inside a neural net to wont reveal how it works, as its reasoning process is embedded in the functions of thousands of simulated neurons, arranged into scores of interconnected layers (Knight, 2017).

Liability: Currently it is not clear exactly how, why and where deployed AI models hold their creators liable and needs to be made crystal clear. For example, in healthcare, models created and trained by third parties are increasingly used by physicians. Where does liability lie? If these models are statistically more accurate than human physicians, will the burden be on physicians or healthcare providers to default to the most accurate solution? Even if they don't understand how the model works or where the data comes from? What if the model makes an error, who is responsible? Does a model train continuously during deployment? If so, the model is reshaping itself based on the data its being exposed to. Who is responsible in that circumstance (Burt, 2018)? What is clear is that at the very least the basic framework for how liability exists in practice needs to be clear before AI model deployment.

"Fail silence": How do we differentiate what is a failure from what isn't? Knowing when to introduce human review and conduct anomaly detection will be crucial. When the input data changes over time, the model continues to make correct decisions, but for reasons that don't make full sense. However, the model may be silently being pushed towards failure, and when the model eventually fails, debugging it becomes an impossible task (Burt, 2018).

An example of this was in 2015, Google's image classifier began classifying black people as gorillas. The engineers had no idea this was going to happen, Google apologized and began to work on the problem. However, three years later, Google was still unable debug and figure out why this was happening. Eventually, Google ended up having to remove

the labels “Gorilla, chimp, chimpanzee, and monkey” from their image classifier to ‘fix’ the problem (Simonite, 2018).

Now while most would argue that an AI system must be able to explain its decisions in human terms in order to be trusted, much of AI’s ingenuity stems from its ability to find data representations that are inherently non-transparent to the human mind.(Danielsson, Macrae, & Uthemann, 2017). For this reason, black box neural networks pose a particularly interesting dilemma:

It seems that there exists a trade-off in some tasks between performance (accuracy) and explainability (Seseri, 2018).

Unencumbered by human cognitive bias, and armed instead with computational brute force, some AI are described as “creative” when they find solutions that humans have not considered, much less attempted to implement. This has been observed many times in videogames, and more famously in the case of AlphaGo’s “Move 37”[6]. The reality is that due to cognitive limitations, humans cannot process all (or even most) of the available information at our disposal when faced with time constraints, and therefore we settle for a satisfactory solution rather than an optimal one. It is this ability to generate unique solutions that makes AI attractive (Scherer, 2015). Some argue therefore, that the level of explainability is always going to be the result of a trade-off, and this needs to be clearly documented (Burt, 2018).

Even if full explainability were possible, this would spawn ‘white-box scenarios’, which as previously discussed, raises cybersecurity concerns about the exploitation of a computer system’s weaknesses. A white-box scenario is where adversarial human agents (attackers) know the internals (architecture and parameters) of the AI system being attacked (Sharif et al., 2016) This would make optimisation against safety-critical AI systems dangerous, and much harder to prevent.

- [6] Google’s AI AlphaGo made a never before seen, “one-in-ten-thousand” move to defeat grandmaster and 18-time world Go champion Lee Sedol in game 2 of a 5-game match that took place in 2016. The infamous move became knowns as “Move 37” (Metz, 2016).

Interrogability: As noted above, different models make different trade-offs between accuracy vs explainability, but on the human side, who can we ask or interrogate if we need to get an accounting for any specific model output? Interrogability is therefore concerned with both the technical and social aspects of explainability (Burt, 2018).

4.6. Autonomous systems / AI type approval

To date, there exists no industry consensus on how to test autonomous /AI systems, particularly in safety-critical environments, as these approaches to computer-based reasoning have been heavily criticized (M. L. Cummings & Britton, 2018). The novelty and unpredictability of autonomous systems means that many failure modes will be unforeseen, and therefore untested and unmanaged until they occur. This scenario has unfortunately been played out in several pedestrian and occupant fatalities involving autonomous vehicles (Skeete, 2018). If reducing the risk of human error is to be the principal benefit of autonomous systems, then autonomous systems must first become more reliable than humans (M. L. Cummings & Britton, 2018).

4.7. AI ethics and principles

The use of military robots, the nature of human-robot relationships (sex partners, caregivers, servants) and the concept of robot/AI personhood, will likely raise ethical questions and concerns in the future. Thus, in addition to establishing regulatory regimes that govern the design and deployment of robots and AI in society, we must consider the need to establish a code of ethics that underpins their operation. Any such value system must reflect a broad normative consensus on what ethical values robots and AI systems must embody (Guihot et al., 2017).

4.8. AI alignment

AI alignment (or value alignment) is the process of ensuring that artificial intelligence systems reliably do what humans want. To achieve AI alignment, we must have a satisfactory definition of human values, gather human value data in a manner compatible with those values, and find reliable ML algorithms that can learn and generalize from this data (OpenAI, 2019).

Alignment gets harder as ML gets smarter for two main reasons:

1. As AI is increasingly applied to consequential and complex tasks (hiring, medicine, scientific analysis, public policy), more reasoning will be required, leading to more complex alignment algorithms.
2. Advanced systems may be capable of answers that sound plausible but are wrong in nonobvious ways. This is not the same an intentional deception. A ML model trained on human data may not be able to differentiate what is "truth" from what humans say is "best". We must therefore recognize that we are unable to answer some types of questions, and therefore, must prevent AI from pretending to answer (OpenAI, 2019).

Unfortunately, the AI goal-alignment problem is currently unsolved, and remains the subject of active research. Training an AI to learn, adopt, and retain our goals are incredibly difficult problems (Tegmark, 2017) and goal system engineering does not yet exist as an established discipline. Furthermore, there is currently known method for transferring human values to a digital computer (value-loading) (Bostrom, 2014). Given the monumental task ahead, it is probably safest to start devoting our best efforts now, to ensure that we'll have the answers when we need them (Tegmark, 2017).

4.9. Recursive Self-Improvement and Superintelligence

It has been argued that AI systems designed to recursively self-improve could lead to a singularity event or 'intelligence explosion' and therefore should be subject to effective safety and control measures. These principles reflect concerns within the industry specifically related to the development of AGI. Developers have considered and even developed various technical contingencies like DeepMind's "big red button" that could be activated in the event that an AI becomes adversarial. The implication is that up to this limit, the "nuclear option" of shutting down a rogue AI completely is always available to its creators. (Guihot et al., 2017).

Unfortunately factors such as controversy, hype and charlatanry may cause 'long term AI concerns' to be shunned by respected scientists and other established figures (Bostrom, 2014). However as stated before, "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014) deserves at least a paragraph or two of consideration here.

Several forms of Superintelligence have been proposed such as Quality Superintelligence which is described as a system that is at least as fast as a human mind but is vastly qualitatively 'smarter' in the performance of anthropocentrically complex cognitive tasks (Bostrom, 2014). However, this paper considers Speed Superintelligence to be a much more relevant concern given all that has been discussed thus far. Speed Superintelligence is an intellect on par with human baseline intellectual capability, but faster (i.e. AGI running on fast hardware. Consider the computational and internal communication advantages of today's modern computer compared with the human brain: Biological neurons operate at a speed of about 200 Hz, while a modern microprocessor performs at 4 GHz. Axons carry action potentials at speeds of 120 m/s or less, whereas electronic processors can communicate optically at the speed of light (300,000,000 m/s). An emulation running at 10,000x that of a biological brain could read a book in a few seconds and write a PhD thesis in an afternoon. At 1 million x an emulation could accomplish an entire millennium of intellectual work in one working day (Bostrom, 2014).

Executing intellection at such speeds would have the effect of dilating time, making human timescales else seem like slow motion in comparison. A speed superintelligence might prefer to work only with digital objects, and commune mainly with other fast minds, rather than with humans, as faster minds face greater opportunity costs in the use (or misuse) of their time (Bostrom, 2014). There is much more to be said on the issue of Superintelligence, however it is beyond the scope of this particular paper.

4.10. Control

Control is the ability for human operators to monitor an AI system and alter its behaviour if necessary. Control is based machine-human communication where a human is still "in the loop" (Tegmark, 2017). The 'Control Problem' (Bostrom, 2014) arises when an AI agent is operating on behalf of a human principal and occurs in the operational phase. Loss of local control (legally responsible humans for operation and supervision) or general control (any human) of AI is a significant concern, especially in the case of the latter. Peter Huber coined the term "public risk" and defined it as threats to human health or safety that are "centrally or mass-produced, broadly distributed, and largely outside the individual risk bearer's direct understanding and control" (Scherer, 2015).

Of course, loss of all human control need not mean significant public risk if the AI is aligned with human values, however human values are currently nearly impossible to define with any precision. Even if its goals were aligned, there is the risk of malignant failure, in particular what Bostrom (2014) calls 'perverse instantiation'. Perverse instantiation is where the AI system discovers a way of achieving its final goal that violates the intentions of the programmers who defined the goal.

An example of perverse instantiation is as follows:

- Final goal: "Make humans smile."
- Perverse instantiation: [Paralyse human facial musculatures into constant beaming smiles] (i.e. manipulating facial nerves).

Needless to say, others (Bostrom, 2014; Tegmark, 2017) have written extensively on the issue of control, preventative measures and the loss of control which far exceed the scope of this paper. However, one of Bostrom's insights in particular is worth sharing here. He argues that 'behaving nicely' while in a box and devising escape plans are both convergent instrumental goals for friendly and adversarial AI's alike. A friendly AI might 'behave nicely' because it defers to humans, and an unfriendly AI might do the same as part of a long-term strategy. An unfriendly AI may try to escape to pursue its own goals, but a friendly AI might do the same if realized that we would be making a mistake by shutting it down and creating another potentially unfriendly AI. Thus, the good behaviour of a system in its juvenile stages fails utterly to predict its behaviour at a more mature stage. But let us return to present times.

One need not accept the possibility of a superintelligence post-apocalyptic scenario to appreciate the control problem. Today's algorithms are capable of autonomously executing commands such as stock trades on timescales measured in nanoseconds, which puts these actions beyond human real-time intervention. Fortunately, stock trades can – and have been – reversed (Scherer, 2015).

5. Conclusion

The following section are various possible concepts and actions that may be worth exploring further with regards to the 'AI problem'.

Useful classifications of AI

A first useful step that could be taken by social scientists and policymakers would be to categorize the various capabilities of AI, maybe a refined version of the basic 'narrow, strong and general intelligence' typology that is currently in use (Guihot et al., 2017). This could be like what was done with categorization of autonomous vehicles Levels 1 – 5. While the current formulation of AV typology has been criticized by some, it remains a useful heuristic for problem solving that all stakeholders can have meaningful discussions around and amend as needed. Some have also suggested that AI could also be broadly categorized by application, e.g. biometrics (fingerprint, iris scanning, voice and facial recognition), safety-critical, etc.

Systems engineering approach

The systems engineering process has been proposed as an ideal, best-case model to guide the major steps of an AI system's development lifecycle. In this process, technical risks are (theoretically) identified early and mitigated through careful design, development, and testing, with a focus on safety, reliability, or redundancy where needed (M. L. Cummings & Britton, 2018).

Engineers are (in theory) able to trace design decisions back to relevant system requirements, to ensure that the end product matches stakeholder needs. Traceability to requirements, therefore, is a key component of the systems engineering process throughout the design and implementation phases. In the United States for example, the Federal Aviation Administration (FAA) has the reputation of approaching regulation in the most conservative and precautionary manner, where new aircraft are extensively reviewed and tested before being allowed to be put to real-world use[7] (M. L. Cummings & Britton, 2018).

[7] Given the level of scrutiny the FAA is known for, it will be interesting to see the final report about on the regulatory failures that resulted in two deadly crashes and the subsequent grounding of the Boeing 737 globally.

The reality is however, that many companies do not fully comprehend the risks embedded within their systems, and the full implications of taking short cuts through the engineering process. Similarly, AI development represents a significant increase in system complexity, where engineers and computer scientists are yet to determine what processes and technologies are required for effective AI and autonomous systems testing and type approval. Given this lack of consensus on testing methods for embedded probabilistic reasoning in AI systems to ensure better-than-human performance, especially in safety-critical applications, it seems prudent that regulators adopt more precautionary measures at this time. (M. L. Cummings & Britton, 2018)

Sector specific regulation

This paper argues that the regulation of AI should be largely kept sector specific, especially at the level of rules and standards. Regulation of medical device software (MDS) for example, is tightly regulated, especially software in 'safety-critical' medical devices (implants, insulin pumps, pacemakers) that have wireless networking protocols. In addition to raising security and safety concerns, safety critical MDS regulation must also address specific issues such as medical data access (e.g. patient and researcher access). Automotive industry regulators on the other hand, once tasked with understanding the intricacies behind catalytic converters, dynamometers, and physical devices must now become much more familiar with complex software that has been integrated into these vehicles (Ohm & Reid, 2016).

Gatekeepers and standards

Some have argued (Bryson & Winfield, 2017; Dafoe, 2018) 'loose coupling' with professional societies and certification agencies would reduce knowledge asymmetries between regulators and firms. This strategy of enrolling gatekeepers has already proven to be successful elsewhere. 'Gatekeepers' are agents or organizations who have a strategic position over those who are the subject of regulation. That position may either be conferred by legal means through the legislature (e.g. auditors, certification agencies, insurers) or by market position (e.g. retailers – tobacco and alcohol age verification, airlines – visas verification).

An example is the ISO, where the vast majority of its standards are certifiable, in a system where the ISO isn't involved in the certification process, or disputes between firms and certifiers. Instead, the ISO uses accreditation and certification to ensure 'correct' implementation of its standards, which is achieved through a system of conformity assessment: firms are certified by an accredited certifier; the certifier is accredited by an accreditor, and accreditors are accredited by ISO. Hence, the ISO's main concern is with the accreditation of the accreditors, which is achieved through ISO inspections, sampling, supplier certification, and peer assessment (Black, 2017).

The benefits of using gatekeepers are that regulators can use these agents as leverage over the regulatees, while the gatekeepers themselves have no particular incentive to subvert the regulatory requirements, as they do not benefit directly from non-compliance. It is acknowledged however that auditors who sign off false accounts likely benefit from continued business with the firm. (Black, 2012)

Risk-based regulation of AI?

Some have suggested that given the high costs and challenges of effective regulatory intervention, regulators should focus on the areas posing the greatest risks, given their often-limited access to resources (Guihot et al., 2017). A risk-based approach usually involves the following:

1. Regulators set the level and types of risks they are willing to tolerate.
2. Regulators conduct risk assessments and estimate the likelihood of risks eventuating.
3. Regulators evaluate these risks, and rank regulated entities accordingly, from low to high levels of risk.
4. Regulators will then allocate resources proportionally to the levels of risk assessed.

Risk-based approaches are usually carried out by the appropriate regulatory agency in consultation with various industry stakeholders. (Guihot et al., 2017)

The allocation of resources and 'regulatory effort' to address risks that are considered most critical has not gone without criticism. Over the past two decades risk-based regulation has been increasingly adopted by regulators in diverse fields such as the

environment, food safety, health and safety, legal services and financial regulation in many OECD countries. The immediate appeal of risk-based regulation is that it organises and prioritises regulatory action, under the premise that that regulators can 'manage uncertainty' (Black, 2012).

However, when designing a risk-based framework, agencies may select and prioritise risks based on incorrect assumptions; they may use wrong indicators and the omit the right ones. Equally, those outside the organisation may contest the prioritisation of risk, though this usually occurs after the risk has crystallised. Those higher up the in the organisation can lose sight and control of those implementing the interventions lower down. Too much time can be spent on risk analysis versus acting in response to them (Black, 2012).

Risk based regulation frequently operates within a political context where regulators require sufficient political licence to operate, irrespective of their formal legal powers. The (now dissolved) Financial Services Authority (FSA)[8] in the UK has stated on several occasions that it would not have had the political support to impose tougher regulation on financial institutions in the boom years preceding the global financial crisis.

It would appear then that risk-based regulation is fundamentally a contradiction: resources are imperfectly allocated, selected and prioritised in the face of unknown and often unknowable risks, with a built-in tolerance for (sometimes predictable) regulatory failure. At the same time however, regulators must convey a sense of control, rationality and equal protection for all, rooted in the logic of governability.

The fate of the FSA is a cautionary tale about the price that a regulator can pay if it is judged (afterwards) to have gotten its risk prioritisation wrong, even if that prioritisation was widely supported by policy makers at the time. Julia Black (2012, p. 1056) subsequently argues that "risk based regulation is thus an inherently contradictory strategy that can never, politically, speak its true name."

[8] The FSA was a quasi-judicial body responsible for the regulation of the financial services industry in the UK between 2001 and 2013. However, due to perceived regulatory failures in the midst of the 2008 GFC, the UK government abolished the FSA in 2013 and split its responsibilities between two new agencies: The Financial Conduct Authority and the Prudential Regulation Authority of the Bank of England.

4.5. Historical regulation of black boxes

While black box neural nets have been sometimes described in this paper as 'inscrutable', the following examples demonstrate how society has mitigated against the negative externalities of black box systems in the past. These may be worth thinking about moving forward.

(US) Equal Credit Opportunity Act (ACOA) 1974: Certain groups faced discrimination in credit scoring decisions, with difficult algorithms that were hard to understand. ACOA was passed to mandate a basic level of transparency by decreasing discrimination and increasing consumer education. As a result of ACOA, credit applicants were entitled to something called "minimum statement of specific reasons", so that they could understand why an adverse credit decision was made. The minimum statement was a list of criteria e.g. bankruptcy, temporary or irregular employment, unable to verify residence, etc.

Now while the statement does not fully break down how a decision is being made, it does give a basic template, making the black box decision less opaque. Transparency therefore does not necessarily lead to increased understanding. Seeing how an algorithm works is not the same as explainability. Even the enforcement document in the above example states that more than four reasons for an adverse credit decision is not actually meaningful, and unlikely to be helpful. Opacity may need to be limited at times, but even then, transparency is not explainability (Burt, 2018).

(US) Supervision and Regulation Letter 11-7 (SR 11-7): SR 11-7 is a 'Supervisory Guidance on Model Risk Management' enforced by the US federal reserve. After the GFC, regulators noticed banks using more complex algorithms, and as a result, banks had less of an understanding of how and why they were making particular decisions. So, SR 11-7 identified two major risks 1) model errors and 2) misused models ("models by their nature are simplifications of reality, and real-world events may prove those specifications inappropriate.") SR 11-7 then proposed a detailed set of guiding principles for managing model risk it called "effective challenge" that is summarised as "critical analysis by informed parties who can identify model limitations and assumptions and produce changes." Hence it is possible to govern, monitor and improve black boxes without understanding them (Burt, 2018).

Agency & Responsibility in humans as a potential model: We treat human decision making in different stages according to age. E.g. parents are initially responsible for their children's actions, then there is an intermediate stage where children become partially responsible for their actions (status of being a minor), and then eventually there is adulthood, where individuals become entirely liable for their actions. Could we therefore classify neural nets in terms of their maturity? Can we use age as a proxy for training data? Type approval for driverless cars in the EU has included calls proposing a 1-million-kilometer drive cycle as the foundation for testing and validation (Skeete, 2018).

However how does the law then deal with our own inability to explain our decisions as adults? One tool is the "Reasonable Person Standard" that is a widely used throughout various areas of the law. This standard asks judges and juries, given all of the data that the person had at the time, given all of the context, did the person act reasonably? It's an incredibly subjective standard and can evolve over time, but subjective standards need not be perfect to be useful in engaging things we do not fully understand. Hence model maturity and subjective standards of reasonableness may be useful (Burt, 2018).

Leveraging recalcitrance

Some researchers (Benthall, 2017) argue that the barriers to AI improvement in performance through algorithmic changes eventually becomes prohibitively high. Said differently, improvements in algorithmic design as a pathway to AGI eventually becomes subject to the law of diminishing returns. They argue that attempting to improve an AI's abilities of prediction would prioritize accessing faster hardware and better data.

The argument follows that the recalcitrance of acquiring faster hardware and better data depends upon – among other things – the availability of resources (i.e. the environment). Thus, if an environment imposes variable search and acquisition costs for hardware and data, then the recalcitrance of these improvements would increase with intelligence[9], which (theoretically) would curtail exponential growth in AI performance (i.e AGI or a superintelligence take-off) (Benthall, 2017).

[9] Rate of change in intelligence (intelligence / time) = Optimisation power / recalcitrance

Thus, if 'intelligence growth' is limited by data and hardware, and not by improvements in artificial intelligence algorithms, Benthall (2017) suggests then, that AI researchers may not be best placed to mitigate the risks of artificial intelligence. He argues rather, that regulators overseeing the use of generic computing hardware and data storage might play a more crucial role in determining the future of artificial intelligence than those designing algorithms (Benthall, 2017).

Might it be feasible and appropriate to monitor AI-relevant resources such as hardware and data? Well-established monitoring regimes already exist for other potentially dangerous technologies, most notably, fissile materials and chemical production facilities for the purpose of implementing nuclear and chemical weapon agreements (Brundage et al., 2018). Can we regulate the rate of intelligence gain? Should intelligence agencies monitor AI groups and projects? Should governments be open to nationalizing of the most advanced AI projects? Should civilian efforts in sensitive areas might be regulated or outlawed? If the goal is to keep track of the most advanced projects, then surveillance based on the best resourced projects may be sufficient (Bostrom, 2014).

This proposition is especially compelling, given that others (Fridman & Brockman, n.d.; Hao, 2019; Scherer, 2015; Sutton, 2019; Tkatch, 2019) have begun to concede "the bitter lesson", which is that increased compute (scale of computational resources) is becoming increasingly significant to attaining state of the art AI performance.

Regtech

Finally, the use of Regtech will be of considerable benefit to authorities, as it would give them the ability to 'optimise the rulebook' and make supervisory processes more robust and cost-effective. AI could scan the literature for new research and advise policy makers of promising new ideas. AI may also be able to replace some applied research, and even take over much of the model writing functions, guided by high-level theories. Ultimately, regulatory AI could provide recommendations to the policy authority, based on its theoretical understanding of the system and provide conditional forecasts of its own behaviour. Such a system will be expected to justify and explain its reasoning (which remains a significant challenge), leaving policy makers to contend with the explainability

dilemma described above, as their first instinct will likely to be reject any advice which cannot be explained to their satisfaction (Danielsson et al., 2017).

Closing remarks

AI has often been described as mankind's last invention, because if we manage to solve it, then in theory we will solve all other problems. Human general intelligence has given our species language, technology, complex social organization and ultimately dominion over the planet. Many other creatures posses greater strength and sharper claws, however it is our human intellection and status as "apex cogitators" (Bostrom, 2014) of this planet that gives us a decisive strategic advantage over all. For now.

Despite some stigmatisation, we must not allow AGI to become a dirty word among the research community by those overly concerned with respectability, as conservative attitudes may ultimately bound and dilute our efforts. As the fate of gorillas now depends on humans, the fate of humans may one day depend on the actions of superintelligent AI, and our only leverage is that we get to build it. Therefore, let us not squander our opportunity as the way to reduce the risks posed by AI today (and tomorrow) will be through strategic analysis and capacity building. Gaining insight and capacity are both elastic endeavours (Bostrom, 2014), where small extra investments today will yield relatively large returns tomorrow, and these returns are compounding, making humanity's subsequent efforts more effective.

References

Allen, D. W. E., & Berg, C. (2018). *Regulation and Technological Change* (SSRN Scholarly Paper No. ID 3140921). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3140921

Arner, D. W., Barberis, J., & Buckley, R. P. (2016a). 150 years of Fintech : An evolutionary analysis. *JASSA*, (3), 22.

Arner, D. W., Barberis, J. N., & Buckley, R. P. (2015). *The Evolution of Fintech: A New Post-Crisis Paradigm?* (SSRN Scholarly Paper No. ID 2676553). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2676553

Arner, D. W., Barberis, J. N., & Buckley, R. P. (2016b). *FinTech, RegTech and the Reconceptualization of Financial Regulation* (SSRN Scholarly Paper No. ID 2847806). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2847806

Benthall, S. (2017). Don't Fear the Reaper: Refuting Bostrom's Superintelligence Argument. *ArXiv.Org*. Retrieved from http://arxiv.org/abs/1702.08495

Berggren, C., Magnusson, T., & Sushandoyo, D. (2015). Transition pathways revisited: Established firms as multi-level actors in the heavy vehicle industry. *Research Policy*, *44*(5), 1017–1028. https://doi.org/10.1016/j.respol.2014.11.009

Black, J. (2012). Paradoxes and Failures: 'New Governance' Techniques and the Financial Crisis. *The Modern Law Review*, *75*(6), 1037–1063. https://doi.org/10.1111/j.1468-2230.2012.00936.x

Black, J. (2017). ‘Says who?’ liquid authority and interpretive control in transnational regulatory regimes. *International Theory*, *9*(2), 286–310. https://doi.org/10.1017/S1752971916000294

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies* (Reprint edition). OUP Oxford.

Brooks, R. (2015). Robotics: Ethics of artificial intelligence. *Nature News*, *521*(7553), 415. https://doi.org/10.1038/521415a

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv:1802.07228 [Cs]*. Retrieved from http://arxiv.org/abs/1802.07228

Bryson, J. J., & Theodorou, A. (2018). *How Society Can Maintain Human-Centric Artificial Intelligence*.

Bryson, J. J., & Winfield, A. (2017). Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer*, *50*(5), 116–119. https://doi.org/10.1109/MC.2017.154

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics*, *130*(3), 453–483. https://doi.org/10.1016/j.jfineco.2018.03.011

Burt, A. (2018). *Regulating Artificial Intelligence: How to Control the Unexplainable*. Retrieved from https://www.youtube.com/watch?v=jcE5JPImxfc&t=1433s

Contag, M., Li, G., Pawlowski, A., Domke, F., Levchenko, K., Holz, T., & Savage, S. (2017). How They Did It: An Analysis of Emission Defeat Devices in Modern

Automobiles. *2017 IEEE Symposium on Security and Privacy (SP)*, 231–250. https://doi.org/10.1109/SP.2017.66

Cummings, M. L. (2019). *Lethal Autonomous Weapons: Meaningful human control or meaningful human certification?* Retrieved from https://hal.pratt.duke.edu/publications

Cummings, M. L., & Britton, D. (2018). *Regulating Safety-Critical Autonomous Systems: Past, Present, and Future Perspectives*. Retrieved from Duke HAL-Humans and Autonomy Lab website: https://hal.pratt.duke.edu/publications

Cummings, M. 'Missy' L. (2017). *Artificial Intelligence and the Future of Warfare*. Retrieved from https://www.chathamhouse.org/publication/artificial-intelligence-and-future-warfare

Dafoe, A. (2018). AI Governance: A Research Agenda. Retrieved June 6, 2019, from The Future of Humanity Institute website: http://www.fhi.ox.ac.uk/

Danielsson, J., Macrae, R., & Uthemann, A. (2017). *Artificial intelligence, financial risk management and systemic risk*. Retrieved from LSE Systemic Risk Centre website: http://www.systemicrisk.ac.uk/publications/SP13

DARPA. (2018). Explainable Artificial Intelligence. Retrieved November 27, 2018, from https://www.darpa.mil/program/explainable-artificial-intelligence

Deloitte. (2017). Driving value creation with technology investments | Deloitte Insights. Retrieved June 2, 2019, from https://www2.deloitte.com/insights/us/en/focus/cio-insider-business-insights/technology-investments-value-creation.html

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., … Song, D. (2017). Robust Physical-World Attacks on Deep Learning Models. *ArXiv:1707.08945 [Cs]*. Retrieved from http://arxiv.org/abs/1707.08945

FCA. (2019, April 16). FCA Research Agenda. Retrieved June 6, 2019, from FCA website: https://www.fca.org.uk/publications/corporate-documents/fca-research-agenda

Foley, S. (2017, March 28). BlackRock cuts ranks of stockpicking fund managers. Retrieved November 27, 2018, from Financial Times website: https://www.ft.com/content/b0056320-13e3-11e7-b0c1-37e417ee6c76

Forbes. (2017). What Companies Are Winning The Race For Artificial Intelligence? Retrieved November 27, 2018, from Forbes website: https://www.forbes.com/sites/quora/2017/02/24/what-companies-are-winning-the-race-for-artificial-intelligence/

Fridman, L., & Brockman, G. (n.d.). *Greg Brockman: OpenAI and AGI | Artificial Intelligence Podcast (MIT AI) - YouTube*. Retrieved from https://www.youtube.com/watch?v=bIrEM2FbOLU

Galloway, S. (2018). *The Four: The Hidden DNA of Amazon, Apple, Facebook and Google*. S.l.: Corgi.

GAO. (1997). *Operation Provide Comfort: Review of U.S. Air Force Investigation of Black Hawk Fratricide Incident*. (OSI-98-4). Retrieved from https://www.gao.gov/products/OSI-98-4

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt Journal of Entertainment and Technology Law*, *20*, 385–456.

Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. Retrieved June 7, 2019, from MIT Technology Review website: https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Knight, W. (2017). There's a big problem with AI: even its creators can't explain how it works. Retrieved November 26, 2018, from MIT Technology Review website: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Lee, I., & Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons*, *61*(1), 35–46.

Lootsma, Y. (2017). Blockchain as the Newest Regtech Application— the Opportunity to Reduce the Burden of KYC for Financial Institutions. Retrieved May 8, 2019, from Initio website: https://www.initio.eu/blog/2017/9/26/blockchain-as-the-newest-regtech-application-the-opportunity-to-reduce-the-burden-of-kyc-for-financial-institutions

Metz, C. (2016, March 16). In Two Moves, AlphaGo and Lee Sedol Redefined the Future. *Wired*. Retrieved from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

Moravec, H. (1998). When Will Computer Hardware Match the Human Brain? *Journal of Evolution and Technology*, *1*(1), 10.

Ohm, P., & Reid, B. E. (2016). *Regulating Software When Everything Has Software* (SSRN Scholarly Paper No. ID 2873751). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2873751

OpenAI. (2019, February 19). AI Safety Needs Social Scientists. Retrieved May 9, 2019, from OpenAI website: https://openai.com/blog/ai-safety-needs-social-scientists/

Opimas. (2017). Opimas: Artificial Intelligence in Capital Markets: The Next Operational Revolution. Retrieved November 27, 2018, from Opimas: We begin with an understanding website: http://www.opimas.com/research/210/detail/

Packin, N. G. (2017). *Regtech, Compliance and Technology Judgment Rule* (SSRN Scholarly Paper No. ID 3043021). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3043021

Rasmussen, R. E. (2007). *The Wrong Target: The Problem of Mistargeting Resulting in Fratricide and Civilian Casualties*. Retrieved from NATIONAL DEFENSE UNIV NORFOLK VA JOINT ADVANCED WARFIGHTING SCHOOL website: https://apps.dtic.mil/docs/citations/ADA468785

Scherer, M. U. (2015). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies* (SSRN Scholarly Paper No. ID 2609777). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2609777

Searls, D. (2019). Blindered by the GDPR | Linux Journal. Retrieved June 9, 2019, from https://www.linuxjournal.com/content/bindered-gdpr

Seseri, R. (2018). The problem with 'explainable AI.' Retrieved June 6, 2019, from TechCrunch website: http://social.techcrunch.com/2018/06/14/the-problem-with-explainable-ai/

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. *Proceedings of*

the 2016 ACM SIGSAC Conference on Computer and Communications Security, 1528–1540. https://doi.org/10.1145/2976749.2978392

Simonite, T. (2018, January 11). When It Comes to Gorillas, Google Photos Remains Blind. *Wired*. Retrieved from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Skeete, J.-P. (2017). Examining the role of policy design and policy interaction in EU automotive emissions performance gaps. *Energy Policy*, *104*, 373–381. https://doi.org/10.1016/j.enpol.2017.02.018

Skeete, J.-P. (2018). Level 5 autonomy: The new face of disruption in road transport. *Technological Forecasting and Social Change*, *134*, 22–34. https://doi.org/10.1016/j.techfore.2018.05.003

Sunstein, C. R. (2003). *Beyond the Precautionary Principle* (SSRN Scholarly Paper No. ID 307098). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=307098

Sutton, R. (2019). The Bitter Lesson. Retrieved June 6, 2019, from http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews*, *39*(1), 103–128. https://doi.org/10.1080/01441647.2018.1494640

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence* (01 edition). Penguin.

Tkatch, D. (2019, April 11). The Very Bitter Lesson. Retrieved June 3, 2019, from Dmitri Tkatch website: https://medium.com/@dmitri.tkatch/the-very-bitter-lesson-b8063a68673e

Transport & Environment. (2019). New EU car emission tests not enough to stop carmakers' cheating | Transport & Environment. Retrieved June 2, 2019, from https://www.transportenvironment.org/press/new-eu-car-emission-tests-not-enough-stop-carmakers%E2%80%99-cheating

Waters, R. (2019, March 12). OpenAI turns itself into a for-profit enterprise. Retrieved June 3, 2019, from Financial Times website: https://www.ft.com/content/3efe0fa6-4438-11e9-b168-96a37d002cd3

Zachariadis, T. (2016). After 'dieselgate': Regulations or economic incentives for a successful environmental policy? *Atmospheric Environment*, *138*, 1–3. https://doi.org/10.1016/j.atmosenv.2016.04.045

Zittrain, J. L. (2005). *The Generative Internet* (SSRN Scholarly Paper No. ID 847124). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=847124