Journal of Artificial Intelligence Research 70 (2021) 1003–1030 Submitted 08/2020; published 03/2021

The Societal Implications of Deep Reinforcement Learning

Jess Whittlestone JLW84@CAM.AC.UK
Leverhulme Centre for the Future of Intelligence
University of Cambridge

Kai Arulkumaran KAILASH.ARULKUMARAN13@IMPERIAL.AC.UK
Imperial College London

Matthew Crosby M.CROSBY@IMPERIAL.AC.UK
Leverhulme Centre for the Future of Intelligence
Imperial College London

Abstract

Deep Reinforcement Learning (DRL) is an avenue of research in Artificial Intelligence (AI) that has received increasing attention within the research community in recent years, and is beginning to show potential for real-world application. DRL is one of the most promising routes towards developing more autonomous AI systems that interact with and take actions in complex real-world environments, and can more flexibly solve a range of problems for which we may not be able to precisely specify a correct 'answer'. This could have substantial implications for people's lives: for example by speeding up automation in various sectors, changing the nature and potential harms of online influence, or introducing new safety risks in physical infrastructure. In this paper, we review recent progress in DRL, discuss how this may introduce novel and pressing issues for society, ethics, and governance, and highlight important avenues for future research to better understand DRL's societal implications.

1. Introduction

Artificial intelligence (AI) is already having an impact on many areas of society (Whittaker et al., 2018; Whittlestone et al., 2019), and further advances in AI research are likely to precipitate much greater impacts. There is a growing community of researchers and practitioners doing excellent work to identify and address ethical and societal issues raised by current applications of AI (Schiff et al., 2020). Given the pace of AI progress, it is crucial that this community also think ahead about what challenges might be raised by future advances, and how forms of governance, norms, and standards being developed around AI today can best prepare us for the future evolution of AI systems. Of course, predicting the future of any technology with certainty is difficult. One way to strike a balance between thinking ahead while still being grounded in reality is to ask the question: what might be the impacts on society if current successful trends in AI research continue and result in new and widespread societal applications?

This paper explores the potential societal implications of one avenue of AI research currently showing promise: Deep Reinforcement Learning (DRL). The paper is primarily aimed at researchers and policy practitioners working on the ethical, societal and governance implications of AI, but may also be of interest to AI researchers concerned with the impacts of

their own work. Our discussion aims to provide important context and a clear starting point for the AI ethics and governance community to begin considering the societal implications of DRL in more depth.

DRL has received increasing attention in recent years, leading to some high-profile successes, particularly in board and video games (Mnih et al., 2015; Silver et al., 2016, 2017; Berner et al., 2019; Vinyals et al., 2019). DRL combines Reinforcement Learning (RL), an approach based on learning through interaction with an environment, with the potential to result in highly autonomous systems (Sutton & Barto, 2018), with Deep Learning (DL), an increasingly popular method that enables AI systems to scale to more complex problems and environments (Goodfellow et al., 2016). Several researchers have suggested that DRL is likely to be a key component of more general-purpose and autonomous AI systems in the future (Arel, 2012; Chen & Liu, 2018; Popova et al., 2018).

Though there are still considerable barriers to large-scale real-world deployment of DRL systems, we are beginning to see potential for application in areas including robotics (Ibarz et al., 2021), online personalisation and targeting (Zhao et al., 2019), finance (Fischer, 2018), autonomous driving (Tai et al., 2016), healthcare (Esteva et al., 2019), and data centre cooling (Gasparik et al., 2018). These applications and others will likely become more widespread as the technology improves. This could have substantial implications for people's lives: for example, by speeding up automation in various sectors, changing the nature and potential harms of online influence, or introducing new safety risks in physical infrastructure. We therefore suggest that now is an ideal time to begin thinking about how more widespread application of DRL-based technologies might impact society, what ethical questions might arise or be made more pressing as a result, how progress on different areas of technical research will affect these ethical challenges, and what forms of governance might be required to mitigate any risks.

One challenge for exploring the potential impacts of future AI progress is that doing so will often require considerable background knowledge about current AI techniques and research directions. Such understanding is often beyond the core expertise of AI ethics and governance scholars, and there is a lack of accessible resources. We therefore begin this paper by providing a broadly accessible introduction to DRL methods with links to further, more in-depth resources for the interested reader.[1]

Following this background, we outline in Section 3 the ways in which plausible future applications of DRL may introduce or exacerbate pressing issues for society, ethics and governance, including by: raising questions for current approaches to human oversight; introducing new challenges for safety and reliability; and changing incentives for data collection. We then discuss how progress on different technical challenges might shape the landscape of these societal concerns: how might breakthroughs in different areas magnify or mitigate societal impacts and risks, or change the direction of the field such that currently underresearched risks become more important? Understanding and monitoring progress on the technical research areas crucial to widespread application of DRL will be key to successful and timely management of future societal impacts, and will require interdisciplinary collaboration between experts in AI, social science, and policy. Based on this analysis, we highlight

1. Readers already familiar with DRL could skip straight to Section 3.

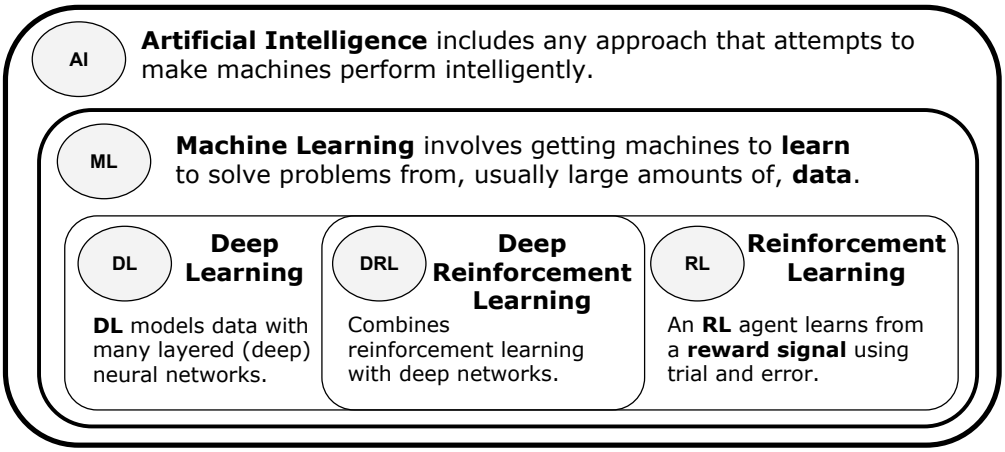

Figure 1: The relationship between deep reinforcement learning and other areas of AI research.

some important avenues for future work and conclude with recommendations for researchers in the AI ethics and governance community.

2. Deep Reinforcement Learning: a Brief Overview

DRL is an approach to AI falling within the subfield of machine learning (ML), which combines two popular approaches in current ML research: reinforcement learning (RL) and deep learning (DL).[2]

In general, the purpose of ML is to find solutions to problems that we cannot directly write programs for. This involves using (often large amounts of) data to optimise the parameters of a mathematical model, which will then encode a solution (Bishop, 2006). This is a powerful approach to solving many different types of problems. For example, while we could easily write a computer program to emulate a medical questionnaire (which is essentially a flowchart), we cannot similarly distill the knowledge of a doctor to diagnose a patient on the basis of a CT (computerised tomography) scan. However, given a large amount of data from CT scans and corresponding expert diagnoses, we can use ML to train a computer program to imitate doctors on this task (Kononenko, 2001; Fatima et al., 2017).

However, many of the problems that we have to solve in the real world require going beyond making predictions based on labelled inputs: we may also want to interact with our *environment* through sets of *actions*, where the choice of action depends on what *goals* we have. Part of how humans learn is through this type of trial and error: when we take an

2. For an excellent and accessible overview of ML also aimed at those interested in its implications, we encourage reading "Machine Learning for Policy Makers" (Buchanan & Miller, 2017). For more in-depth, technical overviews of DRL and its components, DL and RL, we refer readers to the numerous excellent resources already published (Goodfellow et al., 2016; Arulkumaran et al., 2017; Sutton & Barto, 2018; François-Lavet et al., 2018; Kaelbling et al., 1996).

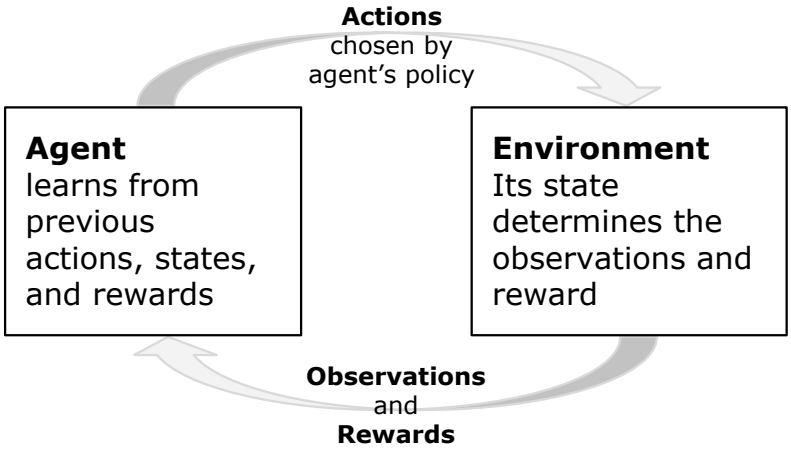

Figure 2: An agent interacts with its environment through its actions, and learns in response to receiving new observations and rewards.

action or make a choice, we get feedback, which gives us some indication of whether we should take similar actions in the future. For example, a young child quickly learns that when they say or do certain things, adults give them positive attention, leading them to repeat those same behaviours for *reward* more often. Reinforcement learning (RL), the core of DRL, is based on this approach to problem-solving and learning: an RL agent[3] learns through interaction with an environment, and, through trial-and-error, can alter its own behaviour in response to the feedback it receives.

More precisely, the goal of RL is to learn a *policy* that recommends the best action to take at any given moment in order to maximise total reward over time. The reward received at any time point is given by a *reward function*, which specifies how good or bad the agent's actions are (but not necessarily exactly what they should be) given the current *state* of the world.

RL takes place in a feedback loop, incorporating perception, action, and learning, as illustrated in Figure 2. An RL agent first perceives the current state of its environment, performs an action in response, and then observes the next state as well as a reward.[4] This loop continues until a predetermined terminal condition is reached, or can run on indefinitely.

A good policy should give the agent an appropriate action to take in every possible state of the environment, including ones the agent has never seen before. For example, when AlphaGo (Silver et al., 2016) beat former world Go champion Lee Sedol, it encountered board positions that had never occurred during its training period, and still played moves at least as good as any human expert. The first generation of RL approaches learned polices in

3. RL systems are often referred to as 'agents', because they act autonomously in their environment. This does not mean that an RL agent is responsible or aware of their actions, just that the human designer is one step removed from the action selection process.
4. In standard RL formulations, the agent gets a reward at every timestep, but that reward may sometimes (or even often) be 0.

the form of large tables, which indicated what action to take in any given state. However, many interesting modern applications of RL have far too many possible states or actions for this tabular approach to be practical or effective. This has led to decades of work on using more complex functions to approximate policies, culminating in recent years in the use of *deep learning* for learning more complex policies.

Deep learning (DL) involves using many-layered models—most commonly artificial neural networks—to represent complex, non-linear functions.[5] A neural network accepts an input (usually a large matrix of numerical values), and, through the repeated application of mathematical operations in each layer, will produce an output (another matrix of numerical values). The trainable parameters of the network, which can number in the billions (Brown et al., 2020), can be updated by a learning algorithm so that, over time, the network ends up producing the desired outputs. Key to the success of DL (LeCun et al., 2015) is its ability to learn high-level features or representations from raw sensory data (e.g., images, text, or audio), which can be used to find patterns in high-dimensional data and generalise to previously unseen inputs.

DRL involves training a neural network using RL, where the network learns which actions to take given its sensory inputs and feedback in the form of rewards.[6] The RL algorithm specifies how to update the parameters of the neural network so that it outputs actions leading to higher rewards. DL has enabled RL to scale to previously intractable problems by improving the performance of RL algorithms in two key ways: through *representation learning*, allowing them to learn directly in complicated environments, and *function approximation*, enabling policies with complex inputs and/or outputs (Arulkumaran et al., 2017).

Because neural networks can approximate any function, they can be used to encode the policy of an agent. This allows RL to extend to problems with state or action spaces that are too large for the tabular approach—which includes most interesting modern applications. Deep networks can also learn features directly from raw sensory data, which means that they can, at least theoretically, be applied in many environments, including the real world—with appropriate sensors. Without deep networks, the inputs to a RL algorithm generally have to be pre-processed and provided by a human expert. By contrast, DeepMind's DQN algorithm—the work that kickstarted the field of DRL—was able to achieve human-level performance on some Atari 2600 games from the game screen and the score, but no additional domain knowledge about any of the games (Mnih et al., 2015).

An important extension of RL is multi-agent RL (MARL), in which multiple agents learn to compete or co-operate in an environment (Busoniu et al., 2008). While some agents may be deployed in isolation, there are many scenarios where groups of agents will need to interact with each other. Some parts of MARL research also look at mixed human-AI interaction (Bard et al., 2020), where understanding the intent of other agents is especially important. As a subfield of RL, MARL inherits all of its technical issues, but also introduces extra emergent complexity. While a full treatment of the intersection of DRL and MARL is

5. DL can involve other types of deep models (Damianou & Lawrence, 2013), but is effectively synonymous with the use of (artificial) neural networks. While artificial neural networks were historically inspired by biological neurons found in the brain, they are far more simplistic, and it is more informative to think of artificial and biological neural networks as independent, loosely similar, systems.
6. Technically speaking the network may output a 'value' for each state and then a further rule will determine the action performed.

out-of-scope for this work, it is worth noting that agents can have different, even conflicting reward functions, and hence a single 'optimal' solution does not exist in the general case, even in theory.

Another implicit assumption in most ML research is that ML models are trained on exactly the same distribution of data that they will be tested on. In the context of DRL, an example would be training a robotic arm to pick and place a certain type of object in a given place in a factory line, 'freezing' the learned model, and then deploying it in the exact same scenario. However, such a model would most likely be unable to adapt to changes in the type of object or placement along the line. To do so, one would also have to incorporate continual learning (Parisi et al., 2019), where the deployed model could continue to learn and adapt to changes in its environment. While trained DRL agents could be deployed without continual learning capabilities, they would be limited to strictly regulated, static environments. As this condition is extremely limiting, we assume that most impactful future applications of DRL will need to be capable of some form of continual learning.

In general, current DRL methods work best on problems that have: (a) a well-defined environment, where it is possible to provide observations of the environment as input, specify what sorts of actions can be taken, and model time as passing in discrete steps; (b) a clear reward function, which is appropriate to the task, provides enough information to learn from, and cannot be easily 'gamed' (Amodei et al., 2016; Krakovna, 2018; Lehman et al., 2020); (c) plenty of data to train the deep neural network—in the context of RL, this means it must be able to interact with and gain feedback from the environment easily and quickly; and (d) lots of computing power available, as exemplified by the high profile successes in DRL (Silver et al., 2016, 2017; Berner et al., 2019; Vinyals et al., 2019). The latter requirement is exacerbated by the fact that ML algorithms also have 'hyperparameters'[7] that need to be tuned separately, requiring the training of many models to find the settings that work best.

In reality, it is rare that all of these ideal conditions are met, which is currently preventing the widespread deployment of DRL systems. However, successful applications of DRL are already apparent, even in these early stages of technological development. These initial applications are already raising issues for society, ethics, and governance which will only become more important as progress is made. We give an overview of these issues in the next section, in particular noting how they relate to current and emerging applications of DRL.

3. Challenges DRL Raises for Society, Ethics, and Governance

Areas where DRL has begun to see practical application include: to send more personally-relevant notifications on Facebook (Gauci et al., 2018); to control Google's data centre cooling system, improving energy efficiency (Gasparik et al., 2018); to automate aspects of electronic trading, such as in J.P. Morgan's LOXM (Bacoyannis et al., 2018); and in robotics and manufacturing (Ackerman, 2020). In the future, we may see more substantial use of DRL in a variety of areas from clinical decision support to resource management, and online education to autonomous driving (François-Lavet et al., 2018).

These potential applications of DRL could have many benefits. More flexible robotic systems could improve manufacturing and delivery processes beyond their current limitations, more adaptive recommender systems could better take account of individuals' personal

7. Hyperparameters are set before learning begins; 'standard' parameters are adjusted during learning.

preferences and longer-term consequences, and improving the efficiency of city infrastructure could help preserve vital and scarce resources. The promise of these benefits ensures that we will see more resources invested in realising these applications in the future. At the same time, the possibility of more autonomous systems that operate flexibly across a range of real-world domains introduces and magnifies a number of problems for AI ethics and governance, which we discuss in this section.

3.1 Human Oversight

The Ethics Guidelines for Trustworthy AI report published in 2019 by the European Commission's High Level Expert Group on AI (High-level expert group on Artificial Intelligence, 2019) states that in order to support human autonomy and decision-making, AI systems must allow for human oversight. Most applications of AI in society today support existing, identifiable decisions taken by individuals and groups, such as doctors, stockbrokers, and lawyers (Parson et al., 2019). This makes human oversight a relatively straightforward matter of ensuring that there is always a human 'in-the-loop', i.e., the ability for a human to oversee the overall activity of the system and intervene in every decision made. Applications of DRL, which aim to increase the autonomy of machines, may pose a challenge for this approach to human oversight.

DRL could potentially be used in ways that go beyond supplementing or even replacing identifiable decisions, towards taking over control of much larger-scale processes, systems and networks, such as those involved in resource management in cities (Mohammadi et al., 2017). What adequate 'human oversight' should look like in these contexts, where a system may be making hundreds of small decisions in a very short period of time, is unclear. The data flow that a system is acting upon may be incomprehensible to humans, or simply too large and fast-moving for meaningful oversight to be maintained. For example, a DRL system designed to monitor and adjust energy usage in a building may be constantly making so many small decisions that it is impractical for a human to review and alter decisions even after the fact, and the system may be sufficiently complex and opaque that it is not possible to identify mistakes and intervene in real-time. It may be possible to impose constraints while the system is being designed—ensuring that the system turns off and a human is alerted if levels go above or beyond a certain threshold, for example. However, there remains a question of whether this constitutes sufficient oversight—or whether in this or similar situations we should avoid deploying DRL systems until ways to maintain more substantive oversight are found.

Another challenge for oversight is introduced by the fact that deployed DRL systems may continue to learn and adapt their behaviour once embedded in their environment, possibly at a pace that is challenging for humans to keep track of and therefore meaningfully oversee. The question of how to maintain human oversight over systems that are continually learning is not one that has been addressed in existing ethics and governance proposals, and will become more pressing as continual learning systems become more advanced and practical. Similarly, if MARL becomes more prevalent, it would make understanding and intervening at the level of a single agent more difficult, posing additional challenges for oversight.

In response to the direction of AI research, the European Commission's more recent whitepaper on AI states that "the appropriate type and degree of human oversight may

vary from one case to another" (European Commission, 2020). Even though it may not be possible for a human to explicitly approve every decision, it may be possible to retain meaningful oversight by ensuring a human is able to review decisions after they have been made, by monitoring an AI system in operation and intervening in real-time if necessary, or by imposing operational constraints in the design phase. Further work on flexible approaches to human oversight should take into account possible evolutions in the autonomy and behaviour of DRL systems discussed here. As systems become more autonomous and promise more high-impact applications, there will inevitably be economic pressures to trust systems to make more decisions and take over more processes, rather than waiting for, or relying on, slow human input. The benefits of doing so need to be traded off against the potential costs to human autonomy that come from reduced oversight, and we must begin thinking about how to navigate this tradeoff now before pressures to implement more autonomous systems grow.

3.2 Safety and Reliability

DRL agents learn by exploring their environment to discover the actions that lead to the highest reward over time. This trial-and-error approach to learning means that in order to learn how to avoid making a mistake, a DRL agent would have to first make that mistake (Berger-Tal et al., 2014). In many real-world contexts this is unacceptable: we cannot have self-driving cars running over pedestrians, or an energy control system accidentally switching off electricity in a hospital, before learning not to do these things. This can cause problems even when DRL is not being deployed in a physical environment: Microsoft's chatbot Tay[8] reproduced thousands of offensive tweets before the problem was noticed and it was taken down. Deploying DRL systems in the real world therefore requires new ways of ensuring the safety and reliability of systems, and in particular requires approaches to safe exploration (Pecka & Svoboda, 2014; Garcıa & Fernández, 2015): how can agents explore enough during training to learn robust behaviour, while avoiding the kinds of exploration that could cause harm?

Ensuring the safety and reliability of DRL systems becomes even more pressing if they are deployed with increasing autonomy in high-stakes domains and systems. For example, it has been suggested that DRL may be able to make effective use of the data generated by 'smart cities' to improve resource management, such as by optimising power usage, increasing the efficiency of traffic signal control systems, and increasing crop productivity by monitoring and correcting parameters in agriculture systems (Mohammadi et al., 2017; Mocanu et al., 2018; Genders & Razavi, 2016). Systems such as energy, transportation, and agriculture are high-stakes in the sense that failures could result in loss of life, and incredibly complex, meaning that their dynamics cannot be fully understood by any single person or group. It is reasonable to suppose that DRL systems, with their adaptability and ability to process huge amounts of data, may ultimately be better equipped to manage these important processes than humans, and improving efficiency could help save critical resources across the world. However, the importance of these domains also means that the risks involved in deploying

8. Although the type of algorithm behind Tay is unknown, this is still a prime example of an agent that interacts with an environment and learns from feedback.

DRL systems are huge, massively raising the stakes of mistakes and unintended consequences compared to smaller-scale and more constrained applications.

Continual monitoring of DRL systems will be essential to ensuring their safety in high-stakes domains, especially if those systems are trained in the real world and continue to learn from a constant stream of data after deployment. Even for systems that are 'frozen' before deployment (i.e., no longer continuing to learn), as initial deployments of DRL likely will be, it is widely accepted that we currently lack adequate methods for verifying safety and reliability (Tarraf et al., 2019; Allen, 2020; Fiander & Blackwood, 2016). For systems that learn over their lifetime, continual monitoring to ensure they do not learn behaviours outside of their intended use is even more challenging (Defense Innovation Board, 2019). Substantial progress on the testing, evaluation, verification and validation of AI systems in general, and DRL systems in particular, will be needed before continually-learning DRL systems can be safely deployed in high-stakes domains.

One research area likely to contribute to this progress is explainable AI (Gunning, 2017; Arrieta et al., 2020). Explainable AI includes the development of models that are inherently easier for humans to understand, but also methods for post-hoc interpretation of standard models such as neural networks. While interpretability methods from the broader field of DL, such as saliency maps (Simonyan et al., 2013) or activation maximisation (Erhan et al., 2009), have been applied to study DRL agents (Greydanus et al., 2018; Such et al., 2018; Dai et al., 2019), the intersection of interpretability and DRL is relatively underexplored. As research in this area matures, policymakers will need to make decisions around acceptable tradeoffs between the fidelity and interpretability of explanations given by AI systems (Ribeiro et al., 2016).

3.3 Harms from Reward Function Design

Beyond more immediate harms, it will also be important to consider the potential longer-term side effects of DRL systems optimising for a specific reward function, which may cause harms which are less obvious but potentially even greater in scale. The flexibility of reward functions—which is what makes DRL systems in theory extremely powerful—introduces much greater potential for unintended consequences, and we might not notice that a system is optimising for slightly the wrong objective before it is too late (Clark & Amodei, 2016; Krakovna, 2018; Lehman et al., 2020). The challenge of reward function design is a long-studied topic in RL and there are many examples where reward functions cause surprising behaviour (Ng et al., 1999; Randløv & Alstrøm, 1998).

There are two related problems for reward function design that can result in unintended harms. First, many real-world tasks do not have easily-defined objectives, and it can be difficult to predict in advance what behaviour a system will exhibit as a result of optimising for a given reward. Second, even if reward functions do correctly capture what the designers or immediate users of a system intended, they may have broader consequences that are harmful to others, perhaps unintentionally.

There are many documented cases of this first problem where DRL algorithms end up learning the 'easiest' way to achieve a stated objective (Clark & Amodei, 2016; Krakovna, 2018). For example, an RL agent trained to play a boat race game was rewarded for getting as many points as possible, under the assumption that this would be equivalent to winning

the race (Clark & Amodei, 2016). However, due to the game layout, the agent was able to find a way to knock over targets by driving in circles, increasing its score but never actually finishing the game. This example is perhaps more amusing than concerning, but such misspecifications could be much more problematic in real-world settings where unintended actions could lead to economic or physical harm. Everitt et al. (2019) give the example of a stock prediction system that predicts the bankruptcy of a company, causing stakeholders to leave. Although it is undesired, creating 'self-fulfilling prophecies' is an effective way for the agent to maximise its predictive accuracy.

To illustrate the second problem, consider the case of social media content-selection algorithms, designed to maximise the likelihood that a user clicks on a given item. Russell (2019) suggests that rather than helping users to find content they are more interested in, ubiquitous use of these algorithms has instead ended up changing users' preferences so that they are more predictable. Worse, this process gradually shifts user preferences towards more extreme content, and therefore contributes to greater conflict online. Russell even goes so far as to suggest the effect of these algorithms has contributed substantially to the resurgence of facism and the crumbling of international bodies like the European Union and NATO.

In large part, the problem here arises as a result of companies optimising for a specific objective at the expense of the interests of individuals and wider society. From the perspective of a social media company, increased engagement can come from people finding more content they genuinely enjoy, or from people being shown more provocative and divisive content. As more companies develop and deploy DRL systems with wide-ranging impacts on users, we must consider both how to ensure that these systems behave as intended over the long-term, and whose interests they are serving. Potential conflicts between commercial and individual needs are particularly concerning here: a reward function by itself may not be sufficient to simultaneously maximise a company's long-term financial return while ensuring there is no negative impact on user agency or welfare, yet it is the former that will likely drive most commercial deployments of AI.

Even with the best of intentions, reward functions are likely to be biased towards those designing or profiting from them. As AI systems come to be deployed more globally, what was once a localised problem becomes a problem of global value alignment: how do we ensure that the goals pursued, both directly and indirectly, by increasingly wide-ranging DRL systems, consider the values of all those impacted by them? Current AI ethics guidelines are predominantly issued by the US and EU, suggesting we have a way to go in ensuring a diversity of perspectives are accounted for (Jobin et al., 2019).

To ensure the longer-term impacts of DRL systems are broadly beneficial, therefore, we need a broader sense of what ethical and responsible reward function design should look like, and how it could be enforced, especially where DRL is being deployed in domains with real implications for people's lives by companies whose goals may not be aligned with society as a whole.

3.4 Incentives for Data Collection

The type and amount of data required to train DRL systems creates incentives for data collection which raise particular ethical concerns. DRL methods tend to require very large

amounts of data as well as ongoing data collection if the DRL system is continuing to learn and adapt its behaviour once deployed. This means that the collection of up-to-date and expansive data from across society could become a key factor for developing and deploying DRL systems. We might therefore be particularly concerned about progress in DRL creating greater incentives for more widespread surveillance. We suggest that particular attention should be paid to advances in continual learning combined with DRL, which would require more real-time data collection, and to advances in methods to learn from large-scale data sources for real-world deployment, such as we are beginning to see in work using DRL to navigate cities using Google street view (Mirowski et al., 2018).

Some areas where DRL might be applied raise greater concerns about data collection and surveillance than others. In order to develop DRL systems for managing infrastructure, for example, it may be particularly important to generate and collect greater amounts of data through the installation of sensors and actuators across cities. This raises a challenging tradeoff between the potentially huge benefits of better resource management against the real threat to people's privacy and autonomy of increasingly 'smart' cities. Another potentially concerning application is the use of DRL to learn more sophisticated and granular models of different users of a digital product or service, and to adapt these models over time based on user behaviour (Zheng et al., 2018). While this could improve people's online experiences, it also might increase the incentive for companies to try to collect or infer increasingly sensitive information about online users, and raises the important ethical question about what data it is acceptable to use to build such models.

There is ongoing technical work in areas such as federated learning and differential privacy attempting to address some of these issues (Ji et al., 2014; Li et al., 2020), and the European Union's General Data Protection Regulation (GDPR) represents an important legislative step towards protecting people's privacy. However, both technical and legislative approaches to data privacy will need to account for how incentives for data collection, and the resulting tradeoffs, may change as DRL systems become more widely deployed.

Increased data collection may also have a disproportionate negative impact on groups who are already vulnerable or discriminated against. Even if surveillance systems are intended for specific, beneficial purposes such as better resource management, the collection of increased data from smart cities consolidates power over individuals' lives in the hands of technology companies and governments, who may well end up using that data for other purposes such as policing and tracking immigrants. Such uses of surveillance have a well-documented history of being used, deliberately or otherwise, to entrench systemic discrimination against minority groups (Crawford et al., 2019). Similarly, increased modelling of users based on online behaviour could easily encode and entrench biased assumptions and stereotypes: Facebook's targeted advertising has been widely criticised for relying on gender and racial stereotypes (Hao, 2019). To the extent that applications of DRL incentivise greater surveillance and enable increasingly sophisticated online targeting, they are likely to exacerbate these problems.

3.5 Security and Potential for Misuse

Without advances in the robustness and security of DRL systems, they may be open to attacks from adversaries attempting to manipulate their performance and behaviour. For

example, even a demonstrably safe DRL-based robot could be forced into dangerous collision scenarios by an adversary perturbing its sensory input or corrupting its reward function. Behzadan and Munir (2018) argue that there are unique challenges posed by ensuring the security of DRL systems. Compared to other ML approaches, it is much less straightforward to distinguish adversarial attacks from benign actions in DRL, since the exploration mechanism means that the training data distribution is continually changing, and delayed rewards may make it difficult to straightforwardly assess the performance of a system. Especially if DRL systems begin to be integrated into critical systems such as healthcare, energy, and transportation, it is essential that key security issues in DRL are well-understood and sufficiently addressed.

There is also a risk that DRL-based systems may be deliberately misused in ways that cause harm. For example, due to their ability to more effectively tailor online content, DRL-based methods could easily be misused to improve the efficacy of online manipulation and disinformation by learning which messages are maximally persuasive to different individuals. In a recent review of online targeting systems, the UK Government's Centre for Data Ethics and Innovation recommend that there should be adequate transparency around online targeting (Centre for Data Ethics and Innovation, 2020). Given its potential use to tailor content more narrowly, it may be particularly important for policymakers and regulators to find ways to mandate transparency around where DRL in particular is used in recommender systems. Improved DRL-based models of human behaviour in particular could be misused to strengthen attempts at social manipulation.

Advances in data efficiency and simulation quality would make it easier for those without access to large amounts of computing power and data to make use of DRL (Tucker et al., 2020). This could make it easier for small groups to misuse DRL capabilities for malicious purposes. For example, Brundage, Avin et al. (2018) describe how cyber attackers might use DRL to "craft attacks that current technical systems and IT professionals are ill-prepared for", emphasising how the feedback loop in RL enables comparison between different approaches to find the one that is most effective at evading security tools. However, DRL could also be used to improve defensive cybersecurity capacities. Whether the advantage lies on the side of the attackers or the defenders is unclear, and would benefit from further analysis, perhaps building on existing work on how technological progress may affect the offense-defense balance(Garfinkel & Dafoe, 2019).

3.6 Automation and the Future of Work

A clear implication of more adaptable and reliable DRL-based systems, especially in robotics and manufacturing, is in changing the susceptibility of different jobs to automation.

Several analyses have suggested that many low-wage or manual jobs are unlikely to be automated in the near future because they require levels of dexterity, mobility and flexibility for which humans still vastly outperform machines (Ford, 2015; Frey et al., 2016; Autor & Dorn, 2013). While this advantage still clearly remains at present, current avenues of progress in robotics led by DRL methods suggest this gap could quickly diminish, a factor that is not considered explicitly by any of the analyses of automation the authors are aware of. Frey and Osborne (2016), for example, discuss areas where AI systems are far from human performance, including in perception and manipulation, and how this limits possibilities for

automation at present, but do not consider the implications of potential progress in these areas. Kai-Fu Lee (2018) argues that manual jobs with the highest risk of replacement are those which are performed in structured environments, with little need for dexterity or social interaction. Prime examples include assembly line inspectors, fruit harvesters and dish washers. However, Lee does not specifically consider how advances in DRL could impact the ability of these jobs to be automated.

Rather than leading to full automation of jobs, advances in DRL may instead ease or improve work in certain domains, by automating tedious or dangerous aspects of manual work. This could have a positive impact on existing employees if it makes their work more enjoyable, but this has to be traded off against potential job losses from improved efficiency. A more detailed analysis of how advances in DRL specifically might speed up automation or change the nature of work in different sectors, for example by improving robotic manipulation, would be valuable.

4. Avenues of Progress in DRL and their Implications

We have argued that widespread deployment of DRL systems in society could introduce new challenges for the ethics and governance of AI, such as by requiring new forms of oversight and new processes for responsible system design, as well as exacerbating the importance of existing challenges, such as ensuring the safety, reliability and security of systems deployed in safety-critical domains. However, current DRL methods are not yet seeing widespread deployment, due to several ongoing technical challenges in DRL research. This gives us time to anticipate and prepare for the societal impacts of DRL. An important part of doing so will involve tracking progress on different technical challenges currently faced by DRL methods, which will enable different kinds of applications. In this section, we therefore introduce some of the key technical challenges for the DRL research community, and discuss the possible implications of progress on these challenges for the ethical, societal, and governance challenges discussed in the previous section.

4.1 Learning in the Real World

DRL systems learn through interacting with an environment. In a simulation or a game the environment can be strictly controlled, but in real-world environments there is more complexity and noise, interaction is often slow and expensive, feedback can be delayed, and performing the wrong action can be dangerous (Dulac-Arnold et al., 2019). For example, in recommender systems the delay between recommending an item to a user and seeing them interact with it may be days or even weeks (Mann et al., 2018), and for robotic agents, a simple fall or crash could damage expensive equipment or even hurt someone. Because of the challenges and risks involved in training DRL in real-world environments, often it is desirable to instead use simulated environments for training. However, good simulations require enormous amounts of data and often still do not result in policies that transfer well to the real world (Pan et al., 2017).

There are two key interrelated approaches to improving DRL systems so that they can be deployed in real-world environments. One approach, which has recently shown impressive results, is finding ways to transfer learned policies from simulation to the real world; these have recently been popularised as 'sim2real' methods. A popular method within these

is 'domain randomisation'—randomising various aspects of the simulation (e.g., lighting, background clutter, friction coefficients, etc.), so that DRL agents learn policies that are robust to variations in these in the real world (Tobin et al., 2017; James et al., 2017; Peng et al., 2018; OpenAI et al., 2019). Other methods include system identification (Chebotar et al., 2019) and domain adaptation (Bousmalis et al., 2018; Tzeng et al., 2020).

Should rapid progress be made in sim2real methods, this would mean that DRL agents could be trained primarily in private simulations, only to be released at the end of successful training. This means that there may be fewer public warning signs before a new system is introduced to the real world, potentially making it more difficult to anticipate potential impacts of new applications in advance, as well as making it easier for companies to develop and deploy systems without oversight and accountability. The resulting systems may be relatively capable, but over-reliance on simulation could also result in the deployment of agents that perform well in simulation but fail catastrophically in the real world. As sim2real methods advance it will be important to consider how risk assessment and oversight of high-risk applications can be ensured.

A second approach to improving the ability of DRL systems to learn in real-world environments is to improve data efficiency, i.e. the ability of DRL systems to learn from much smaller amounts of data. One approach to improving data efficiency in RL is to use model-based RL (MBRL). In MBRL, the agent learns a model of the environment, which can allow the agent to explicitly or implicitly plan what to do without necessitating further interaction with the environment (Sutton, 1990; Hamrick, 2019). In this case, the agent effectively creates its own simulation of the world to learn in. Improvements in model-based RL could also aid with building more realistic simulators from available real-world data. However, substantial progress will be required before it is possible for agents to learn models of most complex real-world environments.

The ability to train DRL systems entirely in simulation or from much smaller amounts of data would reduce the need for 'safe exploration' approaches discussed previously. Mistakes made in simulation would have little or no real cost, and the ability to learn from fewer interactions would significantly reduce (if not entirely mitigate) the risk of serious error. If we see advances in simulated environments then it may also be that the need for extensive real-world data collection is reduced for the purposes of initial training. However, even in such a case it would still be desirable to develop systems capable of continuing to learn once deployed, meaning constant streams of real-world data are still likely to be highly valued. Improvements in data efficiency would therefore also be valuable for reducing incentives for large-scale data collection.

4.2 Safe and Reliable Reward Functions

The behaviour of a DRL agent is highly dependent on its reward function, but as discussed in Section 3, it can be very difficult to specify what we actually want for many real-world tasks. This can make DRL systems unreliable in ways that compromise their safety.

Several different technical approaches are attempting to solve the problem of ensuring the safety and reliability of reward functions. Some approaches focus on ways to train systems without interaction with the real world so that any unintended consequences can be spotted in advance: by training systems 'offline' on batches of historical data (Wiering &

Van Otterlo, 2012; Levine et al., 2020)[9], or in simulated environments as discussed above. Other approaches focus on allowing DRL agents to explore safely by explicitly restricting them from visiting certain states or taking certain actions known or expected to be unsafe, or finding practically viable ways for humans to oversee exploration (Hans et al., 2008; Garcıa & Fernández, 2015; Saunders et al., 2018).

These approaches to ensuring the safety of DRL systems will need to be combined with work on the broader problem of *value alignment*: designing AI systems whose behaviour is in line with human values. One popular approach to solving this problem is *inverse reinforcement learning* (IRL) (Ng et al., 2000; Abbeel & Ng, 2004), where the agent's goal is to infer a reward function from demonstrations, rather than being given one explicitly. This increases the likelihood of the agent learning to behave as intended (especially since it is often easier for us to demonstrate intended behaviour through examples than to provide rules for it). In some approaches, such as Bayesian IRL (Ramachandran & Amir, 2007), the agent also learns a probability distribution over reward functions, meaning learned policies take account of uncertainty, and are likely to be more robust across a wider range of possibilities.

Many real-world tasks also have multi-dimensional objectives, where different subgoals need to be balanced against each other. For example, recommender systems might aim to produce recommendations that achieve some balance of relevance and novelty to the user, among other goals (Aggarwal, 2016). Even for an expert in a domain, it may be difficult to explicitly state what all those different objectives should be or how they should be balanced. An explicit way of dealing with this is multiobjective RL (MORL), where the aim is to either learn a single policy that can adapt to different preferences, or to learn several policies that optimally balance a set of preferences (Liu et al., 2014). Such an approach allows greater flexibility after training, and could be an option for improving interpretability and oversight.

Even when rewards are well-defined, in most settings there is a delay between taking an action and receiving a reward. An agent may have performed thousands of actions immediately prior to receiving a reward and it may be impossible to infer which of those were the key steps that should be repeated in the future. This is known as the credit assignment problem, and is particularly challenging in real-world environments where rewards are sparse (Sutton & Barto, 2018). Current approaches to solving the credit assignment problem try to re-allocate (Arjona-Medina et al., 2019) or augment (Hung et al., 2019) the credit given to past events, but this is still a largely unsolved problem.

If substantial progress is made in any of these methods for safe exploration, value alignment, and/or credit assignment, it will suddenly become possible to deploy DRL systems more safely and effectively in a wider range of real-world environments. These are therefore avenues of progress worth monitoring since they could lead to a very quick proliferation of applications. It should be emphasised that substantial progress in these areas will not be an easy task and is not expected in the near future. Nevertheless, progress is possible, and it is important to ensure that those working on the risks and governance of AI systems are ready for it.

In order to conduct risk assessment for the use of DRL in a robotics system or autonomous vehicle and decide what kind of oversight is needed, for example, one will need to understand what guarantees can be made that the system will not perform certain types

9. In healthcare, offline evaluation is also of particular importance (Gottesman et al., 2018).

of action, and how strict those guarantees are. In order to think through the scope of possible long-term impacts of a new DRL-based information filtering system, one will need to understand how the system balances multiple objectives and accounts for delayed rewards, and how effective those methods will be. As progress is made in all of these areas, therefore, greater communication will be needed between researchers and developers, and those deploying, assessing, and governing AI systems, around what assurances can be made about the behaviour of those systems and the limits of those assurances.

Conversely, if we begin to see increased real-world deployment of DRL systems without seeing considerable progress in these different areas of reward function design, we might be concerned that such systems are being deployed without adequate guarantees that they will behave safely or have the consequences they are expected or intended to.

4.3 Generalisation and Robustness

Ensuring that DRL methods are sufficiently generalisable and robust—that is, they behave as intended across a wide range of scenarios—is a final key challenge for safe real-world deployment. In part, this is a problem inherited from the field of ML more broadly. Most ML algorithms struggle to reuse past knowledge and generalise to new situations (Geirhos et al., 2018), and especially those using deep learning can be frustratingly unstable (Hutson, 2018). DRL has its own approaches to improving generalisation (Justesen et al., 2018; Cobbe et al., 2019; Witty et al., 2018; Zhang et al., 2018b, 2018a), which have strong overlap with broader approaches for generalisation in ML. These include methods in *transfer learning*: developing algorithms that can effectively transfer knowledge from one domain to another (Parisotto et al., 2015; Rusu et al., 2015, 2016), and *meta-learning*, where more generalisable learning methods themselves are learned from data (Schmidhuber, 1987; Wang et al., 2016; Duan et al., 2016; Finn et al., 2017).

A related problem for DRL is that results have proven particularly difficult to reproduce. Henderson et al. (2018) found that results can be significantly impacted by simple changes to the neural network architecture, or by simply multiplying the reward function by a constant value, among other factors. Improving the reproducibility of results may be more a matter of changing research and publication norms, rather than being a research challenge for new methods to overcome. While the ML community in general is particularly in favour of open source material and sharing code, issues of reproducibility still persist and not all research makes its way into the public domain.

Reproducibility problems are exacerbated for real-world applications. Not only do network parameters need to be copied, but also the exact environment in which the agent learns may need to be reproduced faithfully. Until methods become much more robust, seemingly innocent changes in the real-world training environment may have large consequences for the final behaviour of an agent. Further, the expense of replicating real-world settings could be much larger than that of copying and running a simulated environment, further limiting public access to agents for testing and safety analysis.

Without making considerable progress in the areas of generalisation and robustness, it will be difficult to ensure the reliability of DRL systems in real-world domains, which are often highly dynamic and variable. As with reward function design, if we start to see increased use of DRL systems in complex real-world domains without good evidence

of substantial progress in areas such as transfer learning and meta-learning, this should raise concerns about whether such systems might be more brittle than they seem. On the other hand, if we see big breakthroughs in these areas, we may soon see DRL agents capable of performing a wide range of common everyday tasks. This could increase research incentives to solve other related tasks, such as robust real-world navigation and common sense reasoning, which, along with more generalisable DRL methods, could be used to build much more general-purpose systems such as personal assistants and cleaning robots.

5. Discussion

Progress in different areas of technical research will impact the challenges DRL raises for safety, ethics and governance. In general, any progress in research that enables DRL methods to be applied more widely in society will exacerbate these challenges, making them more urgent. For example, substantial progress in the data efficiency of DRL methods and/or the quality of simulated environments for training would likely enable applications in more complex real-world environments, where data is currently either sparse or expensive to acquire. Progress in simulated environments might make DRL for autonomous driving much more feasible, making work on safety and reliability for self-driving cars more pressing (Lin, 2016; Stilgoe, 2018). As a first step, we therefore suggest it would be valuable for those concerned with the wider societal impacts of AI to pay attention to progress in DRL, particularly to anticipate and prepare for the impact of DRL in key domains.

The relationship between technical progress and ethics and governance issues is a little more nuanced than progress always making issues more pressing. In some cases, ethical and societal concerns are being directly addressed by avenues of technical research in DRL, and so progress in these areas could reduce these concerns. While improvements in data efficiency might make certain aspects of DRL safety and reliability more concerning, those same improvements would likely reduce incentives for widespread surveillance, since DRL systems would become less data-hungry. Advances in value alignment—designing reward functions in line with complex, multi-dimensional human values—would reduce concerns about misspecified reward functions having harmful unintended consequences, but might increase the importance of developing standards for reward function design that ensure those methods are used in practice. Improvements in the generalisability of DRL methods could lead to increased confidence that systems will perform reliably across a range of domains, but the resultant increase in applications of DRL could make other issues, such as ensuring that systems can learn safely, more urgent.

It may therefore be particularly important to pay attention to signs of uneven progress across key technical areas in DRL, which might serve as a warning that we could see increased application of DRL without key challenges related to safety, alignment, and security being resolved. For example, if DRL systems begin to be integrated into critical functions in society such as energy or water management, without sufficient assurances about their security, then we might reasonably be concerned about those systems being vulnerable to attack. We need to be able to notice these concerns and work with experts in relevant domains about how to address them before mistakes or accidents happen.

Beyond paying attention to progress in DRL methods and considering their implications, it will also be important for the AI ethics and governance community to explore specific

domains where DRL may be applied in the near future. Throughout our analysis in this paper, we referred to a number of domains where DRL could be applied in the near future, including recommender systems, city infrastructure and manufacturing. Future research could explore the potential impacts of DRL in a variety of important domains not discussed in this paper, such as finance, healthcare, or autonomous vehicles, and attempt to more systematically prioritise which domains of application are likely to raise the most important issues for AI ethics and governance.

While various avenues of technical research are addressing many of the concerns discussed in this paper, one of our key messages is that this will be insufficient by itself. In order to ensure developments in DRL are broadly beneficial, there are many issues and questions that the AI ethics and governance community must also consider. We suggest the two following broad areas will be particularly important as DRL methods move closer to widespread real-world deployment:

1. What kinds of oversight, risk assessment, and technical assurances should we require of DRL systems before deploying them in high-stakes domains, to ensure that they are safe, secure, reliable, and respect human autonomy? How can and should these requirements be implemented practically? How fit-for-purpose are current safety engineering approaches, and what can we learn from the safety engineering community about DRL safety and reliability?

2. How can we ensure that, where DRL systems are deployed with real consequences for individuals' lives, reward functions are designed in ways that take into account the values of all those impacted, and in ways that consider any potential long-term unintended consequences of optimising for a specific objective? What does responsible reward function design look like, and how can it be implemented in practice?

Ideally questions like these would be explored by multidisciplinary teams, combining DRL experts with ethics and governance experts.

6. Summary and Conclusion

DRL is receiving increasing attention in the AI research community, and continued progress could lead to more widespread deployment with profound impacts across many sectors of society. This raises several broad concerns for AI safety, ethics and governance:

- Advances in DRL promise more autonomous systems that can operate effectively with less human oversight, introducing a challenging tradeoff between the potential benefits of such systems and the potential threat to human autonomy and safety;

- DRL systems learn through trial-and-error interaction with an environment, which introduces new challenges for ensuring their safety: we not only need to ensure that systems are safe when deployed, but that they can learn safely;

- More broadly, advances in DRL could enable AI systems to be applied in an increasing number of safety-critical domains, increasing the potential harms of mistakes and therefore increasing the importance of being able to make robust assurances about how systems will behave, and about their security to outside attack;

- The behaviour of a DRL system depends heavily on its reward function, and optimising for slightly the wrong objective can lead to harmful unintended consequences that are difficult to predict in advance—especially if those designing reward functions do not account for the values of all those affected by a system;
- DRL systems currently require large amounts of data to learn effectively, meaning that incentives to deploy real-world DRL applications will likely increase incentives for surveillance and ongoing data collection; and
- Progress in DRL is likely to speed up the susceptibility of jobs to automation in certain areas, such as robotics and manufacturing, which needs to be accounted for in analyses of how AI will impact the future of work.

In order to address these concerns and future challenges, we recommend that the AI ethics and governance community:

- **Find ways to track progress in DRL and its applications**. It is important that those concerned with the wider societal, ethical, and governance implications of AI find ways to track progress in DRL, paying particular attention to progress towards widespread practical deployment. This would help the community better anticipate and prepare for future societal impacts of DRL, and if progress in key areas is uneven, could identify 'warning signs' that DRL risks being deployed in unsafe or harmful ways. This could potentially be an extension of excellent existing initiatives which aim to measure progress in AI more broadly (Perrault et al., 2019; Haynes & Gbedemah, 2019).
- **Consider the implications of DRL progress for existing AI governance initiatives, including standards and regulation**. We have suggested that applications of DRL may introduce new challenges for human oversight and making safety assurances about AI systems, and/or increase the importance of these issues by allowing AI to be applied in more high-stakes domains. A valuable next step would be to consider the practical implications of this for specific governance proposals, which tend to focus on very broad notions of 'AI'. For example, how can guidelines around autonomy in existing AI ethics principles be adapted to apply to DRL systems, which are potentially much more autonomous than any AI systems deployed today? How do notions of safety and security as emphasised in AI ethics proposals need to be adapted to recognise the particular challenges raised by DRL?
- **Establish notions of responsible DRL development**. It could be extremely worthwhile for a multidisciplinary team, combining DRL experts with ethics and governance experts, to explore in more depth many of the questions raised here about how to develop and deploy DRL systems responsibly, especially in high-stakes contexts. What standards of reliability and generalisability should DRL systems be required to meet before deployment, especially where systems are being trained largely in simulation? What does responsible reward function design look like, especially in domains with consequences for individuals' lives, and how can norms of responsibility be embedded in DRL research practices?

DRL is likely to play an important role in the future of AI, promising increasingly autonomous and flexible systems with the potential for application in increasingly high-stakes domains. As a result, DRL introduces and exacerbates a number of concerns which will shape the debate around the safe and responsible use of AI and may be overlooked in more general treatments of AI impacts. Since the most pressing challenges for society will depend on the exact nature of progress of AI research, it will be important to build and maintain strong collaboration between groups working on AI progress and AI governance in order to prepare for the challenges of future AI systems.

Acknowledgements

Correspondence to jlw84@cam.ac.uk. All authors contributed equally.
The authors would like to thank Shahar Avin, Ghazi Ahamat, Miles Brundage, and members of the Kinds of Intelligence Reading Group and AI:FAR Group at the Leverhulme Centre for the Future of Intelligence for helpful discussions and comments. We also would like to thank several anonymous reviewers for their comments, which substantially improved the paper. This work was partly funded by a grant from the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence, and by a grant from the Long-Term Future Fund.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *International Conference on Machine Learning*.

Ackerman, E. (2020). Covariant Uses Simple Robot and Gigantic Neural Net to Automate Warehouse Picking. *IEEE Spectrum Automaton Article*.

Aggarwal, C. C. (2016). An introduction to recommender systems. In *Recommender systems*, pp. 1–28. Springer.

Allen, G. (2020). Understanding AI Technology. Tech. rep., Joint Artificial Intelligence Center (JAIC), The Pentagon United States.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Arel, I. (2012). Deep reinforcement learning as foundation for artificial general intelligence. In *Theoretical Foundations of Artificial General Intelligence*, pp. 89–102. Springer.

Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., & Hochreiter, S. (2019). Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 13544–13555.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, *34*(6), 26–38.

Autor, D. H., & Dorn, D. (2013). How technology wrecks the middle class. *The New York Times*, *24*(2013), 1279–1333.

Bacoyannis, V., Glukhov, V., Jin, T., Kochems, J., & Song, D. R. (2018). Idiosyncrasies and challenges of data driven learning in electronic trading. *arXiv preprint arXiv:1811.09549*.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al. (2020). The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, *280*, 103216.

Behzadan, V., & Munir, A. (2018). The faults in our pi stars: Security issues and open challenges in deep reinforcement learning. *arXiv preprint arXiv:1810.10369*.

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: a multidisciplinary framework. *PloS one*, *9*(4).

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al. (2018). Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *International Conference on Robotics and Automation*, pp. 4243–4250. IEEE.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Buchanan, B., & Miller, T. (2017). *Machine Learning for Policymakers: What It Is and Why It Matters*. Belfer Center for Science and International Affairs.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172.

Centre for Data Ethics and Innovation (2020). Review of online targeting: Final report and recommendations. `https://www.gov.uk/government/publications/cdei-review-of-online-targeting`.

Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., & Fox, D. (2019). Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *International Conference on Robotics and Automation*, pp. 8973–8979. IEEE.

Chen, Z., & Liu, B. (2018). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *12*(3), 1–207.

Clark, J., & Amodei, D. (2016). Faulty Reward Functions in the Wild. `https://blog.openai.com/faulty-reward-functions/`.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., & Schulman, J. (2019). Quantifying generalization in reinforcement learning. *International Conference on Machine Learning*.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., et al. (2019). AI Now 2019 Report. *New York, NY: AI Now Institute*.

Dai, T., Arulkumaran, K., Gerbert, T., Tukra, S., Behbahani, F., & Bharath, A. A. (2019). Analysing deep reinforcement learning agents trained with domain randomisation. *arXiv preprint arXiv:1912.08324*.

Damianou, A., & Lawrence, N. (2013). Deep Gaussian Processes. *Artificial Intelligence and Statistics*, 207–215.

Defense Innovation Board (2019). AI Principles: Recommendations on the ethical use of artificial intelligence by the department of defense. *Supporting document, Defense Innovation Board*.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. Tech. rep. 1341, University of Montreal.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

European Commission (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust. Tech. rep., European Commission.

Everitt, T., Ortega, P. A., Barnes, E., & Legg, S. (2019). Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings. *arXiv preprint arXiv:1902.09980*.

Fatima, M., Pasha, M., et al. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1.

Fiander, S., & Blackwood, N. (2016). House of commons science and technology committee: Robotics and artificial intelligence: Fifth report of session 2016–17. Tech. rep., House of Commons.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 1126–1135.

Fischer, T. G. (2018). Reinforcement learning in financial markets - a survey. Tech. rep., FAU Discussion Papers in Economics.

Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., Pineau, J., et al. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, *11*(3-4), 219–354.

Frey, C. B., Osborne, M., Holmes, C., Rahbari, E., Garlick, R., Friedlander, G., McDonald, G., Curmi, E., Chua, J., Chalif, P., et al. (2016). Technology at work v2.0: The future is not what it used to be. *CityGroup and University of Oxford*.

Garcıa, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, *16*(1), 1437–1480.

Garfinkel, B., & Dafoe, A. (2019). How does the offense-defense balance scale?. *Journal of Strategic Studies*, *42*(6), 736–763.

Gasparik, A., Gamble, C., & Gao, J. (2018). Safety-first AI for autonomous data centre cooling and industrial control. DeepMind blog.

Gauci, J., Conti, E., Liang, Y., Virochsiri, K., He, Y., Kaden, Z., Narayanan, V., Ye, X., Chen, Z., & Fujimoto, S. (2018). Horizon: Facebook's Open Source Applied Reinforcement Learning Platform. *arXiv preprint arXiv:1811.00260*.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 7538–7550.

Genders, W., & Razavi, S. (2016). Using a deep reinforcement learning agent for traffic signal control. *arXiv preprint arXiv:1611.01142*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. (2018). Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*.

Greydanus, S., Koul, A., Dodge, J., & Fern, A. (2018). Visualizing and understanding Atari agents. In *International Conference on Machine Learning*, pp. 1792–1801. PMLR.

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency*, *2*(2).

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16.

Hans, A., Schneegaß, D., Schäfer, A. M., & Udluft, S. (2008). Safe exploration for reinforcement learning.. *European Symposium on Artificial Neural Networks (ESANN)*, 143–148.

Hao, K. (2019). Facebook's ad-serving algorithm discriminates by gender and race. *MIT Technology Review*.

Haynes, A., & Gbedemah, L. (2019). The Global AI Index: Methodology. `https://www.tortoisemedia.com/intelligence/ai/`.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *AAAI*.

High-level expert group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI. Tech. rep., European Commission.

Hung, C.-C., Lillicrap, T., Abramson, J., Wu, Y., Mirza, M., Carnevale, F., Ahuja, A., & Wayne, G. (2019). Optimizing agent behavior over long time scales by transporting value. *Nature Communications*, *10*(1), 1–12.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *American Association for the Advancement of Science*.

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 0278364920987859.

James, S., Davison, A. J., & Johns, E. (2017). Transferring End-to-End Visuomotor Control from Simulation to Real World for a Multi-Stage Task. *CoRL*, 334–343.

Ji, Z., Lipton, Z. C., & Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Justesen, N., Torrado, R. R., Bontrager, P., Khalifa, A., Togelius, J., & Risi, S. (2018). Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89–109.

Krakovna, V. (2018). Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436.

Lee, K.-F. (2018). *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Co., USA.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., et al. (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, *26*(2), 274–306.

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50–60.

Lin, P. (2016). Why ethics matters for autonomous cars. In *Autonomous driving*, pp. 69–85. Springer, Berlin, Heidelberg.

Liu, C., Xu, X., & Hu, D. (2014). Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(3), 385–398.

Mann, T. A., Gowal, S., Jiang, R., Hu, H., Lakshminarayanan, B., & Gyorgy, A. (2018). Learning from delayed outcomes with intermediate observations. *arXiv preprint arXiv:1807.09387*.

Mirowski, P., Grimes, M., Malinowski, M., Hermann, K. M., Anderson, K., Teplyashin, D., Simonyan, K., Zisserman, A., Hadsell, R., et al. (2018). Learning to navigate in cities without a map. *Advances in Neural Information Processing Systems*, 2419–2430.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through Deep Reinforcement Learning. *Nature*, *518*(7540), 529–533.

Mocanu, E., Mocanu, D. C., Nguyen, P. H., Liotta, A., Webber, M. E., Gibescu, M., & Slootweg, J. G. (2018). On-line building energy optimization using deep reinforcement learning. *IEEE transactions on smart grid*, *10*(4), 3698–3708.

Mohammadi, M., Al-Fuqaha, A., Guizani, M., & Oh, J.-S. (2017). Semisupervised deep reinforcement learning in support of IoT and smart city services. *IEEE Internet of Things Journal*, *5*(2), 624–635.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. *International Conference on Machine Learning*.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning.. *International Conference on Machine Learning*, *1*, 663–670.

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., & Zhang, L. (2019). Solving Rubik's Cube with a Robot Hand. *arXiv preprint*.

Pan, X., You, Y., Wang, Z., & Lu, C. (2017). Virtual to real reinforcement learning for autonomous driving. *British Machine Vision Conference*.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71.

Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

Parson, E., Re, R., Solow-Niederman, A., & Zeide, E. (2019). Artificial Intelligence In Strategic Context: An Introduction. `https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/?pdf=321`.

Pecka, M., & Svoboda, T. (2014). Safe exploration techniques for reinforcement learning–an overview. *International Workshop on Modelling and Simulation for Autonomous Systems*, 357–375.

Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. *ICRA*, 1–8.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. C. (2019). The AI Index 2019 Annual Report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*.

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, *4*(7).

Ramachandran, D., & Amir, E. (2007). Bayesian Inverse Reinforcement Learning. *IJCAI*, *7*, 2586–2591.

Randløv, J., & Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. *International Conference on Machine Learning*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Rusu, A. A., Colmenarejo, S. G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., & Hadsell, R. (2015). Policy distillation. *arXiv preprint arXiv:1511.06295*.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Saunders, W., Sastry, G., Stuhlmueller, A., & Evans, O. (2018). Trial without error: Towards safe reinforcement learning via human intervention. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2067–2069.

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 153–158.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. Ph.D. thesis, Technische Universität München.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, *48*(1), 25–56.

Such, F., Madhavan, V., Liu, R., Wang, R., Castro, P., Li, Y., Schubert, L., Bellemare, M. G., Clune, J., & Lehman, J. (2018). An atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. In *NeurIPS Deep RL Workshop*.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine Learning Proceedings*, 216–224.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tai, L., Zhang, J., Liu, M., Boedecker, J., & Burgard, W. (2016). A survey of deep network solutions for learning control in robotics: From reinforcement to imitation. *arXiv preprint arXiv:1612.07139*.

Tarraf, D. C., Shelton, W., Parker, E., Alkire, B., Carew, D. G., Grana, J., Levedahl, A., Leveille, J., Mondschein, J., Ryseff, J., et al. (2019). The department of defense posture for artificial intelligence: Assessment and recommendations. Tech. rep., RAND Arroyo Center, Santa Monica, CA, United States.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 23–30.

Tucker, A. D., Anderljung, M., & Dafoe, A. (2020). Social and Governance Implications of Improved Data Efficiency. *arXiv preprint arXiv:2001.05068*.

Tzeng, E., Devin, C., Hoffman, J., Finn, C., Abbeel, P., Levine, S., Saenko, K., & Darrell, T. (2020). Adapting deep visuomotor representations with weak pairwise constraints. In *Algorithmic Foundations of Robotics XII*, pp. 688–703. Springer.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now 2018 report*. New York: AI Now Institute.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation*.

Wiering, M., & Van Otterlo, M. (2012). Reinforcement learning. *Adaptation, learning, and optimization*, *12*, 3.

Witty, S., Lee, J. K., Tosch, E., Atrey, A., Littman, M., & Jensen, D. (2018). Measuring and characterizing generalization in deep reinforcement learning. *arXiv preprint arXiv:1812.02868*.

Zhang, A., Ballas, N., & Pineau, J. (2018a). A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*.

Zhang, C., Vinyals, O., Munos, R., & Bengio, S. (2018b). A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*.

Zhao, X., Xia, L., Tang, J., & Yin, D. (2019). Deep reinforcement learning for search, recommendation, and online advertising: a survey. *ACM SIGWEB Newsletter*, 1–15.

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., & Li, Z. (2018). Drn: A deep reinforcement learning framework for news recommendation. *World Wide Web Conference*, 167–176.