Defining the Scope of AI Regulations

Jonas Schuett [*]

August 23, 2021

ABSTRACT

The paper argues that policy makers should not use the term artificial intelligence (AI) to define the material scope of AI regulations. The argument is developed by proposing a number of requirements for legal definitions, surveying existing AI definitions, and then discussing the extent to which they meet the proposed requirements. It is shown that existing definitions of AI do not meet the most important requirements for legal definitions. Next, the paper suggests that policy makers should instead deploy a risk-based definition of AI. Rather than using the term AI, they should focus on the specific risks they want to reduce. It is shown that the requirements for legal definitions can be better met by considering the main causes of relevant risks: certain technical approaches (e.g. reinforcement learning), applications (e.g. facial recognition), and capabilities (e.g. the ability to physically interact with the environment). Finally, the paper discusses the extent to which this approach can also be applied to more advanced AI systems.

Keywords: AI regulation, scope of application, legal definition of AI, risk-based regulation

[*] Research Fellow, Legal Priorities Project; jonas.schuett@legalpriorities.org

CONTENTS

1 INTRODUCTION

Policy makers around the world are currently working on AI regulations.[1] The European Commission (2021) recently published a proposal for an *Artificial Intelligence Act*, following their *White Paper on AI* (European Commission, 2020) and the *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on AI, 2019). The US has been more hesitant. Under the Trump administration, the focus was more on removing regulatory barriers (White House, 2019), but this focus is expected to shift under the Biden administration (Engler, 2021). In China, AI regulation is considered a national priority, and the Chinese AI strategy contains explicit goals regarding the development of a regulatory framework (State Council of the People's Republic of China, 2017). This dynamic has already been framed as a 'race to regulate AI' (Smuha, 2021).

One challenge faced by all policy makers who work on AI regulations is how to define the scope of application, which determines whether or not a regulation is applicable in a particular case. The scope of application defines *what* is regulated (material scope), *who* is regulated (personal scope), *where* the regulation applies (territorial scope), and *when* it applies (temporal scope). In this paper, I focus on the material scope. The territorial and temporal scope depend on jurisdiction-specific details, and defining the personal scope is a difficult question which deserves a paper on its own. The scope of application is described in the body of the regulation, using terms typically defined elsewhere in the regulation. These definitions are called legal definitions. The distinction between the terms that are used to define the scope of application ('this regulation applies to AI') and the definitions of these terms ('AI means…') will be important throughout this paper because the core argument is based on the conjunction between the two ('policy makers should only use the term AI for the scope definition if there is a good definition of AI').

Defining the scope of AI regulations is particularly challenging because the term AI is used for so many different systems—'it isn't any one thing' (Stone et al., 2016, p. 48). It can refer to systems that play games (Schrittwieser et al., 2020), produce coherent text (Brown et al., 2020), create fake videos (Korshunov and

[1] By *regulation*, I mean all binding legal rules that are intended to influence the addressees' behavior in order to achieve certain policy objectives in the public interest (Hellgardt, 2016, pp. 48–52). Note that this is not limited to rules created by agencies, which seems to be the classical interpretation of the term by US scholars, and is intended to also include legislation and other binding legal rules more broadly. *AI regulations* are regulations that pursue AI-specific policy objectives, such as reducing AI-specific risks.

Marcel, 2018), predict protein structures (Senior et al., 2020), or diagnose eye diseases (Yim et al., 2020). From a regulatory perspective, these systems have very different risk profiles and therefore must be treated differently. To further complicate things, the term AI is highly ambiguous. There is a vast spectrum of definitions (Legg and Hutter, 2007), and its meaning changes over time. As famously put by John McCarthy: 'as soon as it works, no one calls it AI any more' (Meyer, 2011).

The question of how to define AI in legal terms—especially in a regulatory context—has been raised by many legal scholars. While some have suggested the need for a single legal definition of AI (Lea, 2015; Turner, 2019, pp. 7–8; Martinez, 2019, p. 1022), others have argued that this is not feasible (Reed, 2018, p. 2; Casey and Lemley, 2019, p. 288; Buiten, 2019, p. 45; Gasser and Almeida, 2017). However, there are three notable gaps in the current literature. First, although most arguments rely on certain requirements for legal definitions (e.g. they should be sufficiently flexible to accommodate technical progress), there seems to be no meta-discussion about these requirements. They tend to be treated as something given, without any justification of their legal origin or appropriateness. Second, there is no comprehensive discussion of all requirements; different scholars focus on different requirements. Third, there is only limited discussion of alternative approaches.

The paper proceeds as follows. First, I argue that policy makers should not use the term AI to define the material scope of AI regulations. Next, I argue that policy makers should instead consider using certain technical approaches, applications, and capabilities, following a risk-based approach. Finally, I discuss the extent to which this approach can also be applied to more advanced AI systems.

2 SHOULD POLICY MAKERS USE THE TERM AI TO DEFINE THE MATERIAL SCOPE OF AI REGULATIONS?

The most obvious way to define the material scope of AI regulations would be to use the term AI. For example, the proposed *Artificial Intelligence Act* uses the following formulation:

> This Regulation applies to (a) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country; (b) users of AI systems located within the Union; (c) providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union. (European Commission, 2021, pp. 38–39)

But policy makers should only use the term AI to define the scope of application if they can also define it in a way that is appropriate for regulatory purposes. The question is: does such a definition exist? To answer this question, I propose a set of requirements for legal definitions generally, survey existing AI definitions, and then discuss the extent to which they meet the requirements for legal definitions.

2.1 Requirements for Legal Definitions

In democratic countries, policy makers are bound by higher-ranking sources of law, such as constitutional law and general legal principles. If regulations violate these laws or principles, they can be void or invalid—the particular effects are of course jurisdiction-specific. Here, I give a brief overview of relevant laws and principles in the EU and US and distill them into a list of requirements for legal definitions (Table 1).

Regulations in the EU must comply with the *principle of proportionality*. Pursuant to Article 5(4) of the Treaty on European Union, 'the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties.' Although proportionality has not been used as a general principle of constitutional law in the US, it has nonetheless been recognized as an element of constitutional doctrine in several areas of contemporary constitutional law (Jackson, 2015, p. 3104).

EU regulations must further comply with the *principle of legal certainty*. According to the Court of Justice of the European Union (2009), policy makers are required to ensure 'that Community rules enable those concerned to know precisely the extent of the obligations which are imposed on them. Individuals must be able to ascertain unequivocally what their rights and obligations are and take steps accordingly.'

The US *vagueness doctrine*, which is rooted in due process considerations, has similar implications. According to the US Supreme Court (1926, p. 391), 'a statute which either forbids or requires the doing of an act in terms so vague that men of common intelligence must necessarily guess at its meaning and differ as to its application violates the first essential of due process of law.' Put differently, 'legal protection requires that texts intended in the first place for use by lawyers should be easily understandable by every citizen' (Mattila, 2013, p. 46; see also Price, 2013, p. 1031).

Finally, regulations should be *effective*. Here, effectiveness refers to the degree to which a given regulation achieves or progresses towards its objectives (see European Commission, 2017, pp. 347–348). It is worth noting that the concept of effectiveness is highly controversial within legal research (see De Benedetto, 2018), but for the purposes of this paper, the debate has no relevant implications.

Table 1. Requirements for legal definitions

Title	Description	Origin
Over-inclusiveness	Legal definitions must not be over-inclusive. A definition is over-inclusive if it includes cases which are not in need of regulation according to the regulation's objective (Baldwin et al., 2011, p. 70). Simply put, this is a case of too much regulation.	Principle of proportionality
Under-inclusiveness	Legal definitions must not be under-inclusive. A definition is under-inclusive if cases which should have been included are not included (Baldwin et al., 2011, p. 70). This is a case of too little regulation.	Effectiveness
Precision	Legal definitions must be precise. It must be possible to determine clearly whether or not a particular case falls under the definition.	Principle of legal certainty, vagueness doctrine
Understandability	Legal definitions must be understandable. Ideally, the definition should be based on the existing meaning of terms and comply with the natural use of language. At least in principle, people without expert knowledge should be able to apply the definition.	Principle of legal certainty, vagueness doctrine
Practicability	Legal definitions should be practicable. It should be possible to determine with little effort whether or not a concrete case falls under the definition. The assessment of every element of the definition should be possible on the basis of the information typically available to them.	Good legislative practice (helps to maintain the efficiency of the judicial system)
Flexibility	Legal definitions should be flexible. They should be able to accommodate technical progress. They should only contain elements which are unlikely to change in the foreseeable future.	Good legislative practice (helps to prevent the need for regulatory updating)

To the best of my knowledge, a list similar to Table 1 does not currently exist. Existing lists of requirements for AI definitions (Wang, 2019, pp. 3–6) and scientific definitions in general (Carnap, 1950, p. 7) do not take a legal perspective. And although most of the above mentioned requirements have been discussed in legal scholarship,[2] there seems to be no comprehensive discussion of all requirements.

[2] The problem of over- and under-inclusive AI definitions is discussed by Moses (2007, pp. 260–264), Scherer (2016, pp. 361–362, 373), Reed (2018, p. 2), Martinez (2019, p.

As mentioned above, different scholars focus on different requirements, which tend to be treated as something given and are rarely, if ever, linked to their legal origin.

It is worth noting that the list of requirements should be taken with a grain of salt for two reasons. First, this discussion of the legal origins considers only EU and US laws and principles. Consideration of other jurisdictions was beyond the scope of this paper. However, since the underlying rationale is often not jurisdiction-specific, I expect the list to be useful in other jurisdictions as well. Second, this list is unlikely to be exhaustive. There will likely be further requirements in certain jurisdictions. Similarly, some of the requirements might not be as relevant in some jurisdictions as they are in others, or they might take a slightly different form. For example, it seems plausible that different applications of proportionality analysis (Jackson, 2015) lead to different interpretations of over-inclusiveness. But these variations seem to be a necessary consequence of my attempt to define requirements that are relevant for policy makers worldwide. In any case, the requirements can be used to evaluate existing definitions of AI and can be adapted to the requirements of different jurisdictions.

2.2 Existing Definitions of AI

There is no generally accepted definition of the term AI. Since its first usage by McCarthy et al. (1955), a vast spectrum of definitions has emerged. Below, I provide an overview of existing AI definitions. A more comprehensive collection of definitions can be found in relevant literature (Legg and Hutter, 2007; Samoili et al., 2020). Categorizations of different AI definitions have been proposed by Russell and Norvig (2020), Wang (2019), and Bhatnagar et al. (2018). The OECD (2020) has also presented a framework for the classification of AI systems, which is explicitly targeted at policy makers.

The following list contains popular AI definitions which have been proposed by computer scientists and philosophers:

> The science of making machines do things that would require intelligence if done by men. (Minsky, 1969, p. v)

> The art of creating machines that perform functions that require intelligence when performed by people. (Kurzweil, 1990, p. 14)

1038), Casey and Lemley (2019, pp. 325, 327–328), and Buiten (2019, p. 45). Precision and understandability are addressed by Scherer (2016, p. 373) and Martinez (2019, p. 1035), and flexibility by Moses (2007), Martinez (2019, p. 1017), and Casey and Lemley (2019, p. 357).

> The science and engineering of making intelligent machines, especially intelligent computer programs … Intelligence is the computational part of the ability to achieve goals in the world. (McCarthy, 2007, p. 2)

> That activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. (Nilsson, 2009, p. xiii)

> The study of agents that receive percepts from the environment and perform actions. (Russell and Norvig, 2020, p. vii)

Some legal scholars have also proposed definitions of AI:

> Machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence. (Scherer, 2016, p. 362)

> The ability of a non-natural entity to make choices by an evaluative process. (Turner, 2019, p. 16)

> A system, program, software, or algorithm that acts autonomously to think rationally, think humanely, act rationally, act humanely, make decisions, or provide outputs. (Martinez, 2019, p. 1038)

AI definitions in policy proposals are particularly relevant for this paper:

> Software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. (European Commission, 2021, p. 39)

> (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets. (2) An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action. (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks. (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task. (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting. (Section 238(g) of the FY2019 National Defense Authorization Act; also used by White House, 2020)

> A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. (OECD, 2019a, p. 7)

> The use of digital technology to create systems capable of performing tasks commonly thought to require intelligence. (Office for AI, 2019)

It is worth highlighting a few characteristics of these definitions before continuing with the legal analysis. For example, some of the proposed definitions refer to disciplines ('the science of', 'the art of', 'the study of') and others to systems ('software system', 'artificial system', 'machine-based system'). Most serve academic purposes, while only a few are intended to be used in regulations. One might therefore be tempted to only focus on the definitions by policy makers; however, these definitions are often inspired by academic definitions—for example, the definition in Section 238(g) of the FY2019 National Defense Authorization Act is heavily influenced by Russell and Norvig (2020, pp. 1–5)—thus it seems worthwhile to discuss a wider range of definitions.

2.3 Do Existing AI Definitions Meet the Requirements for Legal Definitions?

As outlined above, legal definitions must meet a number of requirements that can be derived from prior-ranking law, or are at least considered good legislative practice. In Table 2, I discuss the extent to which existing AI definitions meet these requirements using the evaluation options 'Yes', 'No', 'Debatable', and 'Unknown'. Although these options give the false impression that the requirements are binary, they are used for convenience. Since courts ultimately have to make yes-or-no decisions (e.g. whether or not a provision is proportionate), this simplification seems acceptable. It goes without saying that the evaluation is necessarily subjective.

Table 2. Do existing AI definitions meet the requirements for legal definitions?

Requirements	**Existing definitions of AI**
Over-inclusiveness	**No.** Existing AI definitions are highly over-inclusive. For example, many systems that are able to achieve goals in the world are clearly not in need of regulation (e.g. game-playing agents). The same holds true for systems that can, for a given set of human-defined objectives, generate outputs that influence their environment.
Under-inclusiveness	**No.** Some AI definitions are also under-inclusive. For example, systems which do not achieve their goals—like an autonomous vehicle that is unable to reliably identify pedestrians—would be excluded, even though

	they can pose significant risks (Scherer, 2016, p. 362). Similarly, the Turing test (Turing, 1950) excludes systems that do not communicate in natural language, even though such systems may need regulation (e.g. autonomous vehicles).
Precision	**No.** Existing AI definitions are highly vague. Many of them define AI in comparison to human intelligence, even though it is highly disputed how human intelligence should be defined (Legg and Hutter, 2007). Other definitions simply replace one difficult-to-define term ('intelligence') with another ('goal') (Scherer, 2016, p. 361). Russell and Norvig's (2020, pp. 4–5) rational agent definition is equally vague, especially with regards to its notion of limited rationality. In complex environments, agents are often unable to take the optimal action. It is therefore sufficient if they take the action that is optimal *in expectation*. However, in many cases, it is impossible to determine ex-ante whether or not a concrete action is expected to be optimal because ground truth is unattainable. Even if it were, no system can always select the optimal action. How often does a system need to take the optimal action in order to be considered rational?
Understandability	**Debatable.** It is debatable whether existing definitions are understandable. The term seems intuitive at first glance—it is simply a compound of two commonly used terms: 'artificial' and 'intelligence'. However, as mentioned above, it is far from obvious what intelligence actually means. The intuitive meaning may also be misleading. Due to pop-cultural illustrations of AI, people might anthropomorphize AI (Salles et al., 2020; Casey and Lemley, 2019, pp. 353–355).
Practicability	**Debatable.** The practicability of many definitions is also debatable. It may be possible to determine whether or not a system is able to achieve its goals on the basis of typically available information. The Turing test (Turing, 1950), however, would be highly impracticable. Courts would not be able to conduct the test every time they have to decide whether or not a system is considered AI by the law.
Flexibility	**Yes.** The definitions seem sufficiently flexible. The fact that some of them are decades old suggests that they can accommodate technical progress. They also seem relatively general and technology-neutral. One could argue that the so-called 'AI effect' speaks against their flexibility. As McCarthy puts it: 'as soon as it works, no one calls it AI any more' (Meyer, 2011). However, this effect only applies to what is generally considered to be AI. It does not necessarily provide a counterargument against the flexibility of specific definitions.

Taken together, existing definitions of AI do not meet the most important requirements for legal definitions. They are highly over-inclusive and vague, while their understandability and practicability is debatable. I doubt that there even is a definition which meets all of the requirements. I would argue that definitions of the term AI are inherently over-inclusive and vague. Due to its broadness, the term

will always include many different systems with very different risk profiles which must be treated differently.

One might object that this is an inherent property of many legal definitions (Scherer, 2016, p. 373). Many laws use imprecise language, but courts have been able to deal with it. Why should the term AI be any different? My response to this objection is twofold. First, vagueness is a matter of degree. It would be wrong to assume that, simply because courts have been able to deal with imprecise language in the past, policy makers can ignore the issue completely. It might be necessary to use terms that are somewhat imprecise, but I would argue that the term AI is close to the edge of the vagueness spectrum. Second, even if policy makers used a single definition of AI, the above mentioned problems would simply be deferred to the judiciary. Courts would have to develop a casuistry which would also have to meet the requirements detailed above. This would not change the nature of the problem, only the actor who has to solve it.

One might insist that the judiciary would in fact be better suited to develop a precise definition of AI (Casey and Lemley, 2019, pp. 341–344; Turner, 2019, p. 21). I do not argue against this claim, as it seems to be a matter of legal tradition. Scholars from civil law countries (like me) tend to favor statutory definitions, while common law scholars are more used to definitions developed by courts.

Finally, one might point out that the recent proposal for an *Artificial Intelligence Act* does use a single definition of AI (European Commission, 2021, p. 39). Am I really suggesting that the proposal does not meet the requirements for legal definitions? Again, my response would be twofold. First, I would argue that their definition of AI only serves symbolic purposes. The substance lies in Annex I, which contains a list of technical approaches, and Annex III, which contains a list of high-risk applications. In other words, the material scope is only superficially defined by the term AI. Upon closer examination, the term is an 'empty shell', which they have used presumably for communications purposes. Overall, their approach is similar to the one I suggest below. Second, the European Commission was well aware of the above mentioned requirements. The fact that they explain at length why their approach is future-proof (p. 3), proportionate (p. 7), and increases legal certainty (p. 10) suggests that, in their view, other approaches might not meet these requirements.

In summary, the results of my discussion seem defensible against plausible objections. I therefore suggest that policy makers should not use the term AI to define the material scope of AI regulations.

3 WHAT SHOULD THEY DO INSTEAD?

Instead, policy makers should follow a risk-based approach. Risk-based regulation is a regulatory approach that tries to achieve policy objectives by targeting activities that pose the highest risk, which in turn lowers burdens for lower-risk activities (see Black, 2010, pp. 187–190). The scope of such regulations is defined by the risks it wants to reduce. As Turner (2019) puts it, policy makers should not ask 'what is AI?', but 'why do we need to define AI at all?' (p. 8), and 'what is the unique factor of AI that needs regulation?' (p. 15). Or in the words of Casey and Lemley (2019, pp. 342-343): 'We don't need rules that decide whether a car with certain autonomous features is or is not a robot. What we actually need are rules that regulate unsafe driving behavior.'

This approach is in line with existing policy proposals that highlight the importance of risk-based AI regulation. For example, in their proposal for an *Artificial Intelligence Act*, the European Commission (2021, p. 12) focuses on high-risk applications, with almost no requirements for systems with low or minimal risk. They also report that most of the respondents to their stakeholder consultation were explicitly in favour of a risk-based approach (p. 8). Similarly, the German Data Ethics Commission (2019, p. 177) proposes a pyramid of five levels of criticality.

There is an extensive body of literature on risks from AI. Risks have been conceptualized as accident risks (Amodei et al., 2016), misuse risks (Brundage et al., 2018), and structural risks (Zwetsloot and Dafoe, 2019). One could also distinguish between near-term and long-term risks, but some scholars have argued convincingly that this distinction is not always useful, mainly because many ethics and safety issues span different time horizons (Baum, 2018; Cave and Ó hÉigeartaigh, 2019; Prunkl and Whittlestone, 2020).

There has also been some work on AI risk factors, broadly defined as all factors that contribute to risks from AI. Most notably, Hernández-Orallo et al. (2019) have conducted a survey of known safety-relevant characteristics of AI. They distinguish between (1) internal characteristics (e.g. interpretability), (2) effect of the external environment on the system (e.g. the ability of the operator to intervene during operation), and (3) effect of the system on the external environment (e.g. whether the system influences a safety-critical setting).

Although their categorization is convincing, I do not use it below, mainly because it serves a different purpose. Theirs is intended to reveal neglected areas of research and to suggest design choices for reducing certain safety concerns, whereas I am interested in defining AI risk factors in a way that meets the requirements for legal definitions. Their categorization also excludes risks caused

by 'the malicious or careless use of a correctly-functioning system' (p. 1), which would be relevant in a regulatory context. For similar reasons, I also do not use the categorization by Burden and Hernández-Orallo (2020).

Instead, I use my own simple categorization of AI risk factors. I distinguish between (1) technical approaches ('how it is made'), (2) applications ('what it is used for'), and (3) capabilities ('what it can do'). In the following, I explain each of the three categories along with examples and discuss the extent to which they meet the requirements for legal definitions.

3.1 Technical Approaches

Some AI risks are directly linked to certain technical approaches. One such approach is *reinforcement learning*, which is used in games (Schrittwieser et al., 2020), robotics (OpenAI et al., 2018), and recommender systems (Afsar et al., 2021). But using this approach poses a number of inherent risks. For example, if the objective function of a reinforcement learning agent contains explicit specifications only regarding the main goal, it might implicitly express indifference towards other aspects of the environment. This can lead to situations where the agent disturbs its environment in negative ways while pursuing its main goal. This problem is typically referred to as 'negative side effects' (Amodei et al., 2016, pp. 4–7). Another problem is 'reward hacking', the exploitation of unintended loopholes in the reward function (Clark and Amodei, 2016). A third problem is how we can ensure that agents can be safely interrupted at any time (Orseau and Armstrong, 2016). Policy makers who want to address these risks could use the following definition:

> 'Reinforcement learning' means the machine learning task of learning a policy from reward signals that maximizes a value function. (Sutton and Barto, 2018, p. 6)

Policy makers could also use the terms *supervised learning* and *unsupervised learning* to define the material scope of AI regulations. These approaches are used in a wide range of different systems, including systems that support judicial decision-making (Angwin et al., 2016) or select employees (Dastin, 2018). However, both approaches can lead to discrimination by reproducing biases contained in the training data (Bolukbasi et al., 2016; Buolamwini and Gebru, 2018). They can be defined as follows:

> 'Supervised learning' means the machine learning task of learning a function that maps from an input to an output based on labeled input-output pairs. (Russell and Norvig, 2020, pp. 652–653)

> 'Unsupervised learning' means the machine learning task of learning patterns in an input even though no explicit feedback is supplied. (Russell and Norvig, 2020, pp. 652–653)

3.2 Applications

Other risks are not linked to technical approaches, but certain applications. For example, policy makers may want to reduce the risks that *autonomous driving* poses to road safety and security, physical integrity, and property rights. The material scope of such regulations could be defined using six levels of automation, as described in the technical standard 'SAE J3016' (SAE International, 2021).These definitions have already been adopted by policy makers in the US (US Department of Transportation, 2018) and the EU (European Commission, 2018).

Policy makers may also want to reduce the specific risks of *facial recognition technology*. A number of studies show that facial recognition technology can have gender or race biases (Buolamwini and Gebru, 2018). This is particularly worrying if such systems are used for law enforcement purposes. In the US, some municipalities have therefore started to ban state use of facial recognition technology for law enforcement purposes, including San Francisco (Conger et al., 2019) and Boston (Johnson, 2020). The European Commission (2021) has proposed a similar ban in the EU, with a few narrow exceptions. In addition to discrimination risks, facial recognition also raises severe privacy concerns (Erkin et al., 2009). Policy makers who want to address these risks could use the following definition:

> 'Facial recognition' means the automatic processing of digital images which contain the faces of individuals for identification, authentication/verification or categorisation of those individuals. (Article 29 Data Protection Working Party, 2012, p. 2)

3.3 Capabilities

A third category of AI risk factors ist a system's capabilities. For example, policy makers may want to limit the material scope to systems which can *physically interact with their environment* via robotic hands (OpenAI et al., 2019) or other actuators. Only embodied systems can directly cause physical harm or damage property (Hernández-Orallo et al., 2019, p. 2). This ability could be defined as follows:

> 'Physical interaction' means the ability to use sensors to perceive the physical environment and effectors to manipulate this environment.

Another capability-related risk factor is the *ability to make automated decisions*. This would exclude systems which only make suggestions while humans make the final decision. One could call systems with this ability 'self-executive'. Policy makers could use this element to address certain risks resulting from a loss of control (Scherer, 2016, pp. 366–369) and other assurance risks—those risks which stem from an operator's inability to understand and control AI systems during operation (Ortega and Maini, 2018). This element is already being used in Articles 13(2)(f), 14(2)(g) and 15(1)(h) of the GDPR. It can be defined as follows:

> 'Automated decision-making' means the ability to make decisions by technological means without human involvement. (Article 29 Data Protection Working Party, 2018, p. 8)

A third example of a capability is the *ability to make decisions which have a legal or similarly significant effect*. Consider two virtual assistants: one reminds you on your friends' birthdays, the other is able to buy products. Clearly, the two systems have very different risk profiles (the latter may require some degree of consumer protection, for example). This element is already being used in Article 22 of the GDPR. The European Data Protection Board has endorsed the definition by the Article 29 Data Protection Working Party (2018, pp. 21–22):

> 'Legal effect' means any impact on a person's legal status or their legal rights.
>
> 'Similarly significant effect' means any equivalent impact on a person's circumstances, behavior or choices. This may include their financial circumstances, access to health services, employment opportunities or access to education.

3.4 Do Definitions of Certain Technical Approaches, Applications, and Capabilities Meet the Requirements for Legal Definitions?

Let us now examine to what extent definitions of certain technical approaches, applications, and capabilities meet the requirements for legal definitions. Table 3 breaks down the discussion by category and requirement.

Table 3. Do definitions of certain technical approaches, applications, and capabilities meet the requirements for legal definitions?

Requirements	Technical approaches	Applications	Capabilities
Over-inclusiveness	**No.** There will always be systems that use one of the above mentioned technical	**Yes.** In many cases, the main regulatory goal will be to reduce	**No.** Not all systems with certain capabilities pose risks

	approaches, but should not be subject to regulation (e.g. game-playing agents based on reinforcement learning).	certain application-specific risks (e.g. discriminatory recommender systems used to support judicial decision-making).	which are in need of regulation. For example, industrial robots and vending machines both have the ability to physically manipulate their environment, but their risk profile is very different.
Under-inclusiveness	**No.** Relevant risks can not be attributed to a single technical approach. For example, supervised learning is not inherently risky. And if a definition lists many technical approaches, it would likely be over-inclusive.	**No.** Not all systems that are applied in a specific context pose the same risks. Many of the risks also depend on the technical approach.	**No.** Relevant risks can not be attributed to a certain capability alone. By its very nature, capabilities need to be combined with other elements ('capability of something').
Precision	**Yes.** It is easy to determine whether or not a system is based on a certain technical approach.	**Yes.** Applications can be defined precisely. This is by no means a novel challenge for the law.	**Yes.** In many cases, capabilities can be defined in a binary way (e.g. a system either can physically manipulate its environment or not).
Under-standability	**Yes.** For developers it will be easy to understand definitions of certain technical approaches. One can expect the same from non-technical people who are responsible for the development, deployment, or use of systems.	**Yes.** There are no apparent reasons for why definitions of applications are not understandable.	**Yes.** Most capabilities are intuitive (e.g. the ability to physically manipulate its environment).
Practicability	**Yes.** The required information about the technical approach is easy to obtain.	**Yes.** The required information about the application is easy to obtain.	**Yes.** Some capabilities already have established legal definitions (e.g. the ability to make decisions which have a legal or similarly significant effect).
Flexibility	**Unknown.** It is highly uncertain whether today's	**Debatable.** While some applications	**Yes.** Definitions of capabilities seem to be

technical approaches will be used in the future. Definitions will be more flexible if the technical approach is defined broadly, but they will also be less precise.	are unlikely to change in the future, almost certainly new applications will emerge.	able to accommodate technical progress.

In summary, definitions of certain technical approaches, applications, and capabilities meet more of the requirements for legal definition than definitions of the term AI (see Table 2). This suggests that policy makers should favor a risk-based approach over the 'classical' approach.

One might be tempted to simply pick one of three categories, but I would argue that a *multi-element approach* seems preferable (Casey and Lemley, 2019, p. 356). The following example illustrates the idea:

> This regulation applies to facial recognition systems for law enforcement purposes based on supervised learning.

In the example, the material scope is defined by a certain application (facial recognition for law enforcement purposes) and a certain technical approach (supervised learning). This approach allows policy makers to target risks in a more fine grained way and thereby reduce over-inclusiveness and increase precision.

The European Commission's (2021) proposal follows a similar approach. As mentioned above, the material scope is not really defined by the term AI. Instead, the scope definition combines a number of technical approaches (Annex I) with certain high-risk applications (Annex III). (The third element—the ability to generate outputs that influence environments—seems to not play any meaningful role.) Although this seems like a reasonable approach, I would point out three potential areas for improvement. First, to the extent that my observation is correct, the European Commission should consider making it explicit that their scope definition does not rely on the term AI (e.g. in the recitals). This could help to prevent misconceptions among laypeople (e.g. the false interpretation that the regulation would apply to any use of Bayesian statistics, as implied by Carpenter, 2021). Second, they should consider distinguishing between different technical approaches. In the current version, it is sufficient if a system is based on any of the technical approaches listed in Annex I. However, a recruiting system based on a simple statistical approach would not pose the same risks as a system based on supervised learning. Third, they should consider defining capabilities, as doing so could further reduce over-inclusiveness and increase precision.

4 CAN THIS APPROACH ALSO BE APPLIED TO AGI REGULATIONS?

Future AI systems that achieve or exceed human performance in a wide range of cognitive tasks have been referred to as 'artificial general intelligence (AGI)' (Goertzel and Pennachin, 2007). Even though the prospect of AGI is speculative, and some people remain sceptical (Etzioni, 2016; Mitchell, 2021), a number of surveys show that many AI researchers do take it seriously (Baum et al., 2011; Müller and Bostrom, 2016; Grace et al., 2018).

While the development of AGI could be overwhelmingly beneficial for humanity, it could also pose significant risks. Potential risks from AGI have been studied, among others, by Bostrom (2014), Dafoe (2018), Russell (2019), and Ord (2020). There are also a number of public figures, such as Stephen Hawking (Cellan-Jones, 2014), Elon Musk (Gibbs, 2014), and Bill Gates (Rawlinson, 2015), who have warned against the dangers of AGI. Against this background, it is not surprising that policy makers have started taking AGI more seriously (OECD, 2019b, p. 22; White House, 2020, p. 2).

If and when it becomes evident that AGI is in fact possible, policy makers may want to reduce the associated risks via regulation. This would again raise the question of how they should define the material scope of such AGI regulations. Would a risk-based approach be applicable to define all sorts of AI, including AGI?

It seems very likely that the *technical approach* that is used to build AGI will significantly influence its risks and potential risk mitigation strategies. For example, if AGI is developed using reinforcement learning (Silver et al., 2021), we might use an approach called 'reward modelling' to align it to human values (Leike et al., 2018). One might therefore be tempted to rely on technical approaches when defining the material scope of AGI regulations. However, there is an ongoing debate about whether today's technical approaches are sufficient to build AGI. While some AI researchers think this is reasonable (Christiano, 2016), others remain sceptical (Kokotajlo, 2019). Given the high degree of uncertainty, policy makers should probably not rely exclusively on specific technical approaches.

Since AGI is characterized by the generality of its intelligence, it seems less fruitful to define specific *applications*. However, one could nonetheless distinguish between different types of AGI, such as question-answering, command-executing, or non-goal-directed systems (Bostrom, 2014, pp. 177–193). Since these types could influence the feasibility and desirability of different safety precautions (Bostrom, 2014, pp. 191–192), policy makers may want to use them to define the material scope of AGI regulations.

As mentioned above, the decisive *capability* of AGI is the generality of its intelligence. If a system exceeds human intelligence across the board, humanity

would become the second most intelligent species on Earth (Ngo, 2020) and might permanently lose its influence over the future (Bostrom, 2014; Ord, 2020). However, I doubt that there is a definition of this capability that meets the requirements for legal definitions, mainly because I expect it to be highly vague. Instead, policy makers may want to define capabilities that could lead to the development of AGI. These capabilities seem easier to define, but would still capture relevant AGI risks. One such capability could be the ability to recursively self-improve (Bostrom, 2014, p. 409; Good, 1966, p. 33):

> 'Recursive self-improvement' means an agent's ability to iteratively improve its own performance.

In summary, it seems plausible that policy makers could follow a risk-based approach to define the material scope of AGI regulations, though the focus might shift from technical approaches to capabilities.

5 CONCLUSION

In this paper, I have shown that existing definitions of AI do not meet the most important requirements for legal definitions. Therefore, policy makers should not use the term AI to define the material scope of AI. I have also shown that definitions of the main causes of relevant risks—certain technical approaches, applications, and capabilities—meet more of the requirements for legal definitions than definitions of the term AI. Finally, I have argued that this approach can, in principle, also be used to define the material scope of AGI regulations.

The paper has made four main contributions. First, it has provided a comprehensive legal argument for why policy makers should not use the term AI for regulatory purposes and why a risk-based definition of AI would be preferable. Second, it has proposed a list of specific requirements for legal definitions which can also be used to evaluate other definitions. Third, the paper has suggested a new categorization of the main causes of AI risks that policy makers may want to address. And fourth, it can be seen as a first step towards AGI safety regulation, which I expect will turn into its own field of interest for policy makers and researchers in the future.

The findings of this paper are relevant for policy makers worldwide. They support the European Commission's (2021) risk-based approach. The suggested definitions of certain technical approaches, applications, and capabilities can also be used to amend or substantiate the list of techniques and approaches in Annex I and high-risk applications in Annex III. But I expect the findings to be even more relevant for policy makers who have not yet drafted concrete proposals. Defining

the material scope of AI regulations requires careful consideration. I hope this paper comes at the right time to help policy makers rise to this challenge.

ACKNOWLEDGEMENTS

I am grateful for valuable comments and feedback from Seth Baum, Conor Griffin, Renan Araújo, Leonie Koessler, Nick Hollman, Suzanne Van Arsdale, Markus Anderljung, Matthijs Maas, and Sébastien Krier. I also thank the participants of a seminar hosted by the Legal Priorities Project in February 2021. All remaining errors are my own.

REFERENCES

Afsar, M. M., Crump, T. and Far, B. (2021) 'Reinforcement Learning Based Recommender Systems: A Survey', *arXiv preprint arXiv:2101.06286*.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016) 'Concrete Problems in AI Safety', *arXiv preprint arXiv:1606.06565*.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine Bias [online]. *ProPublica*. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 19 July 2021]

Article 29 Data Protection Working Party (2012) 'Opinion 02/2012 on Facial Recognition in Online and Mobile Services', WP192.

Article 29 Data Protection Working Party (2018) 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679', WP251rev.01.

Baldwin, R., Cave, M. and Lodge, M. (2011) *Understanding Regulation: Theory, Strategy, and Practice*. 2nd ed. Oxford: Oxford University Press.

Baum, S. D. (2018) 'Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence', *AI & Society*, 33, pp. 565–572.

Baum, S. D., Goertzel, B. and Goertzel, T. G. (2011) 'How Long Until Human-Level AI? Results from an Expert Assessment', *Technological Forecasting and Social Change*, 78 (1), pp. 185–195.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M. et al. (2018) 'Mapping Intelligence: Requirements and Possibilities', in V. C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence*. Berlin: Springer, pp. 117–135.

Black, J. (2010) 'Risk-Based Regulation: Choices, Practices and Lessons Being Learnt', in *OECD Reviews of Regulatory Reform: Risk and Regulatory Policy*, pp. 185–236.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. and Kalai, A. (2016) 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings', in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4356–4364.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. 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. et al. (2018) 'Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', *arXiv preprint arXiv:1802.07228*.

Buiten, M. C. (2019) 'Towards Intelligent Regulation of Artificial Intelligence', *European Journal of Risk Regulation*, 10 (1), pp. 41–59.

Buolamwini, J. and Gebru, T. (2018) 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Burden, J. and Hernández-Orallo, J. (2020) 'Exploring AI Safety in Degrees: Generality, Capability and Control', in *Proceedings of the Workshop on Artificial Intelligence Safety co-located with 34th AAAI Conference on Artificial Intelligence*, pp. 36–40.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Carpenter, B. (2021) EU Proposing to Regulate the Use of Bayesian Estimation [online] *Statistical Modeling, Causal Inference, and Social Science*. Available from: https://statmodeling.stat.columbia.edu/2021/04/22/eu-proposing-to-regulate-the-use-of-bayesian-estimation [Accessed 21 July 2021].

Casey, B. and Lemley, M. A. (2019) 'You Might be a Robot', *Cornell Law Review*, 105, pp. 287–362.

Cave, S. and Ó hÉigeartaigh, S. S. (2019) 'Bridging Near- and Long-Term Concerns About AI', *Nature Machine Intelligence*, 1 (1), pp. 5–6.

Cellan-Jones, R. (2014) Stephen Hawking Warns Artificial Intelligence Could End Mankind [online]. *BBC*. Available from: https://www.bbc.com/news/technology-30290540 [Accessed 13 July 2021].

Christiano, P. (2016) Prosaic AI Alignment [online]. Available from: https://ai-alignment.com/prosaic-ai-control-b959644d79c2 [Accessed 13 July 2021].

Clark, J. and Amodei, D. (2016) Faulty Reward Functions in the Wild [online]. *OpenAI*. Available from: https://openai.com/blog/faulty-reward-functions [Accessed 19 July 2021].

Conger, K. Fausset, R. and Kovaleski, S. F. (2019) San Francisco Bans Facial Recognition Technology [online]. *The New York Times*. Available from: https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html [Accessed 19 July 2021].

Court of Justice of the European Union (2009) 'Gottfried Heinrich', Case C-345/06.

Dafoe, A. (2018) 'AI Governance: A Research Agenda', Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. Available from: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf [Accessed 13 July 2021].

Dastin, J. (2018) Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women [online]. *Reuters*. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-in...-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Accessed 15 July 2021].

De Benedetto, M. (2018) 'Effective Law from a Regulatory and Administrative Law Perspective', *European Journal of Risk Regulation*, 9 (3), pp. 391–415.

Engler, A. (2021) 6 Developments That Will Define AI Governance in 2021 [online]. *Brookings Institution*. Available from: https://www.brookings.edu/research/6-developments-that-will-define-ai-governance-in-2021 [Accessed 13 July 2021].

Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I. and Toft, T. (2009) 'Privacy-Preserving Face Recognition', in *Proceedings of the 9th International Symposium on Privacy Enhancing Technologies*, pp. 235–253.

Etzioni, O. (2016) No, the Experts Don't Think Superintelligent AI is a Threat to Humanity [online]. *MIT Technology Review*. Available from: https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity [Accessed 13 July 2021].

European Commission (2017) 'Better Regulation "Toolbox"'. Available from: https://ec.europa.eu/info/sites/info/files/better-regulation-toolbox.pdf [Accessed 13 July 2021].

European Commission (2018) 'On the Road to Automated Mobility: An EU Strategy for Mobility of the Future', COM(2018) 283 final.

European Commission (2020) 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust', COM(2020) 65 final.

European Commission (2021) 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)', COM(2021) 206 final.

Gasser, U. and Almeida, V. A. F. (2017) 'A Layered Model for AI Governance', *IEEE Internet Computing*, 21 (6), pp. 58–62.

German Data Ethics Commission (2019) 'Opinion of the Data Ethics Commission'. Available from: https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf [Accessed 13 July 2021].

Gibbs, S. (2014) Elon Musk: Artificial Intelligence is our Biggest Existential Threat [online]. *The Guardian*. Available from: https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat [Accessed 13 July 2021].

Goertzel, B. and Pennachin, C. (eds.) (2007) *Artificial General Intelligence*. Berlin: Springer.

Good, I. J. (1966) 'Speculations Concerning the First Ultraintelligent Machine', *Advances in Computers*, 6, pp. 31–88.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2018) 'When Will AI Exceed Human Performance? Evidence from AI Experts', *Journal of Artificial Intelligence Research*, 62, pp. 729–754.

Hellgardt, A. (2016) *Regulierung und Privatrecht*. Tübingen: Mohr Siebeck.

Hernández-Orallo, J., Martínez-Plumed, F., Avin, S. and Ó hÉigeartaigh, S. S. (2019) 'Surveying Safety-relevant AI Characteristics', in *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019*, pp. 57–65.

High-Level Expert Group on AI (2019) 'Ethics Guidelines for Trustworthy AI'. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed 13 July 2021].

Jackson, V. C. (2015) 'Constitutional Law in an Age of Proportionality', *Yale Law Journal*, 124 (8), pp. 2680–3203.

Johnson, K. (2020) Boston Bans Facial Recognition Due to Concern About Racial Bias [online]. *VentureBeat*. Available from: https://venturebeat.com/2020/06/24/boston-bans-facial-recognition-due-to-concern-about-racial-bias [Accessed 19 July 2021].

Kokotajlo, D. (2019) A Dilemma for Prosaic AI Alignment [online]. *AI Alignment Forum*. Available from: https://www.alignmentforum.org/posts/jYdAxH8BarPT4fqnb/a-dilemma-for-prosaic-ai-alignment [Accessed 13 July 2021].

Korshunov, P. and Marcel, S. (2018) 'DeepFakes: A New Threat to Face Recognition? Assessment and Detection', *arXiv preprint arXiv:1812.08685*.

Kurzweil, R. (1990) *The Age of Intelligent Machines*. Cambridge, MA: MIT Press.

Lea, G. (2015) Why We Need a Legal Definition of Artificial Intelligence [online]. *The Conversation*. Available from: https://theconversation.com/why-we-need-a-legal-definition-of-artificial-intelligence-46796 [Accessed 13 July 2021].

Legg, S. and Hutter, M. (2007) 'A Collection of Definitions of Intelligence', in *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, pp. 17–24.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V. and Legg, S. (2018) 'Scalable Agent Alignment via Reward Modeling: A Research Direction', *arXiv preprint arXiv:1811.07871*.

Martinez, R. (2019) 'Artificial Intelligence: Distinguishing between Types & Definitions', *Nevada Law Journal*, 19 (3), pp. 1015–1042.

Mattila, H. E. S. (2013) *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas*. 2nd ed. New York, NY: Routledge.

McCarthy, J. (2007) What is Artificial Intelligence? [online]. Available from: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf [Accessed 13 July 2021].

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. (1955) 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence'. Available from: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf [Accessed 13 July 2021].

Meyer, B. (2011) John McCarthy [online]. *Communications of the ACM*. Available from: https://cacm.acm.org/blogs/blog-cacm/138907-johnmccarthy/fulltext [Accessed 13 July 2021].

Minsky, M. (1969) *Semantic Information Processing*. Cambridge, MA: MIT Press.

Mitchell, M. (2021) 'Why AI is Harder than We Think', *arXiv preprint arXiv:2104.12871*.

Moses, L. B. (2007) 'Recurring Dilemmas: The Law's Race to Keep up with Technological Change', *Journal of Law, Technology & Policy*, 2, pp. 239–285.

Müller, V. C. and Bostrom, N. (2016) 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion', in V. C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*. Cham: Springer, pp. 555–572.

Ngo, R. (2020) AGI Safety from First Principles [online]. *AI Alignment Forum*. Available from: https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ [Accessed 13 July 2021].

Nilsson, N. J. (2009) *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York, NY: Cambridge University Press.

OECD (2019a) 'Recommendation of the Council on Artificial Intelligence', OECD/LEGAL/0449.

OECD (2019b) 'Artificial Intelligence in Society'. Available from: https://doi.org/10.1787/eedfee77-en [Accessed 13 July 2021].

OECD (2020) 'A First Look at the OECD's Framework for the Classification of AI Systems, Designed to Give Policymakers Clarity'. Available from: https://oecd.ai/wonk/a-first-look-at-the-oecds-framework-for-the-classification-of-ai-systems-for-policymakers [Accessed 13 July 2021].

Office for AI (2019) 'A Guide to Using Artificial Intelligence in the Public Sector'. Available from: https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector [Accessed 13 July 2021].

OpenAI et al. (2018) 'Learning Dexterous In-Hand Manipulation', *arXiv preprint arXiv:1808.00177*.

OpenAI et al. (2019) 'Solving Rubik's Cube with a Robot Hand', *arXiv preprint arXiv:1910.07113*.

Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity*. New York, NY: Hachette Books.

Orseau, L. and Armstrong, S. (2016) 'Safely Interruptible Agents', *Machine Intelligence Research Institute*. Available from: https://intelligence.org/files/Interruptibility.pdf [Accessed 20 July 2021].

Ortega, P. A. and Maini, V. (2018) Building Safe Artificial Intelligence: Specification, Robustness, and Assurance [online]. *Medium*. Available from: https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1 [Accessed 19 July 2021].

Price, J. (2013) 'Wagging, Not Barking: Statutory Definitions', *Cleveland State Law Review*, 60, pp. 999–1055.

Prunkl, C. and Whittlestone, J. (2020) 'Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 138–143.

Rawlinson, K. (2015) Microsoft's Bill Gates Insists AI is a Threat [online]. *BBC*. Available from: https://www.bbc.com/news/31047780 [Accessed 13 July 2021].

Reed, C. (2018) 'How Should We Regulate Artificial Intelligence?', *Philosophical Transactions of the Royal Society A*, 376.

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Penguin Random House.

Russell, S. and Norvig, P. (2020) *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson Education.

SAE International (2021) 'Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles', J3016_202104.

Salles, A., Evers, K. and Farisco, M. (2020) 'Anthropomorphism in AI', *AJOB Neuroscience*, 11 (2), pp. 88–95.

Samoili, S., López-Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F. and Delipetrev, B. (2020) 'AI Watch: Defining Artificial Intelligence', *European Commission*. Available from: https://doi.org/10.2760/382730 [Accessed 20 July 2021].

Scherer, M. U. (2016) 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies', *Harvard Journal of Law & Technology*, 29 (2), pp. 353–400.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S. et al. (2020) 'Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model', *Nature*, 588, pp. 604–609.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T. et al. (2020) 'Improved Protein Structure Prediction Using Potentials from Deep Learning', *Nature*, 577, pp. 706–710

Silver, D., Singh, S., Precup, D. and Sutton, R. S. (2021) 'Reward is Enough', *Artificial Intelligence*, 299.

Smuha, N. A. (2021) 'From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence', *Law, Innovation and Technology*, 13 (1), pp. 57–84.

State Council of the People's Republic of China (2017) 'New Generation of Artificial Intelligence Development Plan'. Available from: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017 [Accessed 13 July 2021].

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G. et al. (2016) 'Artificial Intelligence and Life in 2030', Stanford University. Available from: http://ai100.stanford.edu/2016-report [Accessed 16 July 2021].

Sutton, R. S. and Barto, A. G. (2018) *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press.

Turing, A. M. (1950) 'Computing Machinery and Intelligence', *Mind*, 59 (236), pp. 433–460.

Turner, J. (2019) *Robot Rules: Regulating Artificial Intelligence*. Cham: Palgrave Macmillan.

US Department of Transportation (2018) 'Preparing for the Future of Transportation'. Available from: https://www.transportation.gov/av/3/preparing-future-transportation-automated-vehicles-3 [Accessed 20 July 2021].

US Supreme Court (1926) 'Connally v. General Construction Co.', 269 U.S. 385.

Wang, P. (2019) 'On Defining Artificial Intelligence', *Journal of Artificial General Intelligence*, 10 (2).

White House (2019) 'Maintaining American Leadership in Artificial Intelligence', Executive Order 13859.

White House (2020) 'Guidance for Regulation of Artificial Intelligence Applications'. Available from: https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf [Accessed 13 July 2021].

Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C. et al. (2020) 'Predicting Conversion to Wet Age-Related Macular Degeneration Using Deep Learning', *Nature Medicine*, 26, pp. 892–899.

Zwetsloot R. and Dafoe, A. (2019) Thinking About Risks from AI: Accidents, Misuse and Structure [online]. *Lawfare*. Available from: https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure [Accessed 13 July 2021].