big data and cognitive computing

MDPI

Article

AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk

Brandon Perry [1,*] and Risto Uuk [2]

[1] Independent Researcher, Berkeley, CA 94709, USA
[2] Effective Altruism Estonia, Tallinn 12618, Estonia; ristouuk@gmail.com
* Correspondence: brandonperryofficial@gmail.com

Received: 5 April 2019; Accepted: 2 May 2019; Published: 8 May 2019

Abstract: This essay argues that a new subfield of AI governance should be explored that examines the policy-making process and its implications for AI governance. A growing number of researchers have begun working on the question of how to mitigate the catastrophic risks of transformative artificial intelligence, including what policies states should adopt. However, this essay identifies a preceding, meta-level problem of how the space of possible policies is affected by the politics and administrative mechanisms of how those policies are created and implemented. This creates a new set of key considerations for the field of AI governance and should influence the action of future policymakers. This essay examines some of the theories of the policymaking process, how they compare to current work in AI governance, and their implications for the field at large and ends by identifying areas of future research.

Keywords: policymaking process; AI risk; typologies of AI policy; AI governance

1. Introduction

Artificial intelligence, especially artificial general intelligence (AGI), has the ability to dramatically impact the future of humanity [1]. Notable researchers, such as Bostrom (2014), have expressed concern that advanced forms of artificial intelligence, if not aligned to humans values and wellbeing, could be potentially disastrous and pose an existential threat to our civilization [2]. The two main branches of research on risk from advanced AI are AI safety, which seeks to ensure that advanced AI is engineered in such a way that it will not pose a threat; and AI governance, which focuses on political and social dynamics (AI macrostrategy) and forecasting timelines for AI development [3]. Issues that AI governance looks at include arms race dynamics, social and economic inequality, public perceptions, issues in surveillance, and more.

There has been a modest amount of work on developing policy solutions to AI risk, with a recent literature review by Baum (2017) [4] and Everitt (2016) [5] covering most of it. Some authors have focused on the development of AGI, with proposed solutions ranging from Joy (2000) [6] who calls for a complete moratorium on AGI research, to Hibbard (2002) [7] and Hughes (2007) [8], who advocate for regulatory regimes to prevent the emergence of harmful AGI, to McGinnis (2010), who advocates for the US to steeply accelerate friendly AGI research [9]. Everitt et al. (2017) [5] suggests that there should be an increase in AI safety funding. Scherer (2016) [10], however, at least in the context of narrow AI, argues that tort law and the existing legal structures, along with the concentration of AI R&D in large visible corporations like Google, will provide some incentives for the safe development of AI. Guihot et al. (2017) [11] also notes that attempts to future-proof laws tend to fail, and pre-emptive bans and regulation tend to hurt the long-term health of the field, instead arguing for a soft-law approach. Other authors have focused on the community of researchers, with Baum (2017) [12] promoting a

social psychology approach to promote community self-regulation and activism, and Yampolskiy and Fox (2013) [13] advocating for review boards at universities and other research organizations.

Some authors have advocated for an international approach to resolving AI risk. Erdelyi and Goldsmith (2018) [14] advocated for an international soft-law regime that would serve as a "international forum for discussion and engage in international standard setting activities". Erdelyi and Goldsmith's proposal, however, is not targeted towards AGI risk, although they could scale up to AGI. Wilson (2013) [15] and Bostrom (2014) [2], on the other hand, call for some form of international agreement or control on AGI R&D, with the former advocating specifically for a treaty.

These approaches are necessary given some of the risks, including states pursuing AGI for unprecedented military and economic strength with destabilizing effects (Shulman 2009) [16], and the concentration of wealth and political influence in large corporations (Goertzel 2017) [17]. Questions regarding whether or not AGI R&D should be open sourced or not have been explored by Goertzel (2017) [17] and Bostrom (2017) [18]. Shulman (2009) [16] and Dewey (2015) [19] follow a different approach and advocate for a global surveillance regime to monitor for rogue AGI projects, with Goertzel (2012) [20] suggesting that a limited form of AGI could do this.

As far as current and future research goes, the Future of Humanity Institute has developed an extensive research agenda [3] for AI governance, with three main research areas: Technical landscape, which seeks to understand what artificial intelligence can do and its limits; AI politics, which looks at the political dynamics between firms, governments, publics, etc.; and ideal governance, which looks at possible ways and arrangements for stakeholders to cooperate. This research agenda highlights key issues such as security challenges, international political dynamics and distribution of wealth, and arms race dynamics. Other researchers have published reports dealing with issues such as dual use, similarity, and possible interactions with the cybersecurity community [21] the role and limits of principles for AI ethics [22], justice and equity [23], and AGI R&D community norms [5].

Thus far, much of the literature on AI risk has discussed policy issues, but few studies have talked about how policies are made or how the dynamics of the policymaking process affect their work. Calo (2017) [23] touches upon the problem, noting that there is a lack of institutional expertise, policy tools, and flawed mental models of what AI is, which plague governments' abilities to regulate AI. Scherer (2016) [10] cites certain aspects of the technology itself, such as its ability to be created without special equipment, as a hindrance to the ability to regulate it. Everitt et al. (2017) [5] also briefly discusses policy and political dynamics in the context of AGI researchers, suggesting that AGI researchers should work with other organizations to mitigate the negative dynamics of framing AGI development as an arms race [24]. Finally, the Future of Humanity Institute's research agenda for AI governance [3] touches on policymaking in a few ways, noting that public opinion can have major impacts on technology policy and governance schemes can be subject to mission drift and asking how to facilitate the transition from the present state of affairs to our ideal vision for the future.

This paper continues along the lines of facilitating the transition from the present state to "our ideal vision" by exploring the missing discussion on the role of policymaking in AI governance. Research thus far has largely focused on what problems are out there and what should be done to fix them. However, this paper does not only argue that proposal implementation that takes into account the features of the 'policymaking cycle' may be vital to success in reducing AI risk but that this model actually has massive implications for the research field as a whole. Proposals will be much more effective if they are informed by an understanding of the political and administrative considerations of consensus-building and implementation and could make the difference between making an impact or none at all.

The goal of this paper is to attempt to create a clearer launching point for discussions on the key considerations of the policymaking process for AI governance and the political considerations underpinning policy solutions for AI risk. The policymaking process includes: Problem identification/agenda setting, policy formulation, policy adoption, implementation, and evaluation. Each step of the policymaking process will have different aspects that are critical for the creation of public policies that are able to

effectively reduce AI risk. Each section covers a brief overview of the literature, assesses its implications for the greater AI governance field, and identifies different points where further research is needed. The papers we selected are the primary sources of these different theories of the policymaking process.

The first section maps out and defines terms in the field of AI governance, to give readers a better understanding of how our paper contributes to the way AI governance is approached. We also created a typology for AI risk policies, to provide an understanding as to how AI governance has implications in a diverse range of policy communities and how that interplays with strategic considerations. The next section goes through each step of the policymaking cycle, with a basic overview of some of the literature and discussing its implications for AI governance. It should be noted that the literature covered in each field is not extensive, and further research may be necessary. The last sections cover some of the key implications and limitations.

2. Terms and Definitions

On a broad level, the question of mitigating AI risk, or risks that stem from the development and use of artificial intelligence (such as global catastrophic risks from misaligned AI or military instability from adopting new types of weapons), is broken down into AI technical safety and AI governance. AI technical safety focuses on solving computer science problems around issues like misalignment and the control problem for AGI [2]. AI governance, on the other hand, studies how humanity can best navigate the transition to advanced AI systems [3]. This would include the political, military, economic, governance, and ethical considerations and aspects of the problem that advanced AI has on society.

AI governance can be further broken down into other components, namely the technical landscape (how technical developments depends on inputs and constraints and affects rates or domains of capability improvement), ideal governance (what would we do ideally if we could cooperate), and AI politics (how AI will affect domestic politics, political economy, international relations, etc.) [3]. From these research areas, the problems and solutions necessary to discuss AI policy can be defined. This paper, however, refers to this as AI risk policy to differentiate policies intended to reduce catastrophic risk to society versus policies that apply to AI in any other circumstances.

Policies, however, must be implemented into the legal statutes of government in order to work. Flynn (2017) [25], in the blog post that defines 'AI strategy' [3], also defines 'AI policy implementation', which is carrying out the activities necessary to safely navigate the transition to advanced AI systems. This definition implies it is action-oriented work done in government, policy, lobbying, funding, etc. As mentioned in the endnotes of Flynn (2017), however, there is an implicit gap between AI strategy (governance) research and policy implementation, with no AI policy research that identifies mechanisms for actualizing change.

However, there is another gap that this paper intends to address, which is that the processes that create and implement policies (the policymaking process) often either distort the original policy, fall short of, or even work counter to the intended outcome, or render certain policy options unactionable. Similarly, The AI governance: A Research Agenda report has neither this consideration nor a definition of policy implementation. This paper intends to put forth a definition of AI policymaking strategy to fill this gap, which is defined as:

AI Policymaking Strategy: A research field that analyzes the policymaking process and draws implications for policy design, advocacy, organizational strategy, and AI governance as a whole.

This goes further than the concern listed in the endnotes and also develops an upstream approach to AI governance, where work in implementation in turn feeds back and can provide new insights to AI governance research.

AI policymaking strategy would fit under the definition of AI governance and would be its own subfield in the same way technical landscape is and would help to clarify questions and considerations in the other subfields. AI politics and ideal governance seem to ask questions about what risks humanity faces and what it ought to do about them, approaching the world as if from above and

making corrections, whereas policymaking strategy asks questions about how and what can be done, given both present and future circumstances, and the methods to do so at hand. They approach the world as agents who individually influence the trajectory of the world. These two groups, when they work together, should ideally converge on a policy program that both works and is pragmatic—constituting of policies that both aim at the correct goals and can actually get there.

An example of this would be the proposed solution by Goertzel (2012) [20] of creating a surveillance artificial narrow intelligence that monitors the world to prevent the development of superintelligence. Let us say that Policy X is written to do this. However, Policy X, like all other policies, is not simply just a solution to the problem but a set of intended actions and procedures taken by the government that must first be passed by government [26]. This begs three questions: Can this policy realistically be implemented by government? How do policymakers ensure that Policy X results in the intended outputs and outcomes? And how can policymakers create policy and advocacy strategies to increase the chances of both of these happening? For example, while Policy X is intended to install a surveillance apparatus to prevent superintelligence, would Policy X still have that output and outcome after going through the legislature and executive branch? Is there a chance over time that it would result in mission creep? Policymakers can also develop strategies to ensure that Policy X has its intended outcomes, such as oversight mechanisms within the policy itself. Policymakers can go a step further and ask how the policymaking process itself creates implications for the AI governance field. For example, are there restrictions within the policymaking process that impact timelines for reducing risk, such as how fast governments can act or create new laws? Could some form of upstream innovation be acheived where the policymaking process inspires or generates new ideas for AI governance [27]?

3. Typologies of AI Policy

Before this paper can delve into the policymaking process, AI policy needs to be further refined to understand what kind of policies are being made. The point of this section is to show that AI risk policies are not monolithic, but rather there are multiple approaches to help achieve the same goal, and each set of these policies is going to have with it a different set of political difficulties. It also begs the question in terms of AI governance as a whole as to which sets of policies should be implemented and when, and which policies should be considered relevant to AI risk. In the same way that Bostrom (2014) [2] argues that there may be a preferred order of technological development, there is a similar analog with AI risk policies where there is a strategic order to policies that should be attempted to be implemented, whether it is because their political-capital cost is lower, the cost of failure is lower, or because it helps with future efforts to implement policies (such as the creation of an advisory body).

A typology of AI policies already has some previous explorative work to build on. Brundage (2016) [28] proposed the idea of De Facto AI policies. These are policies that already exist and are relevant to AI. These are further broken down into direct, indirect, and relevant policies. Direct policies are policies that specifically target AI, such as regulations on self-driving cars. Indirect policies are policies that do not specifically target AI but generally impact the development and diffusion of technologies (including AI), such as intellectual property laws and tort law. Relevant policies do not immediately impact AI but are still worth considering because of their impact, such as education policy or the use of electronic medical records.

Brundage (2016) [27] in this paper, however, does not talk about AI risk policy but rather existing policies around AI as a whole. However, the classification used in this paper is useful overall and can be extended into AI risk policy. Instead of whether or not it directly or indirectly affects AI, AI risk policy can be classified into whether or not it directly or indirectly aims at reducing AI risk. Direct AI risk policies would explicitly govern the use, development, deployment, etc. of AI to reduce risk. Examples of direct AI risk policy could include funding for AI safety research, rules for the development of AGI, international agreements on AI, etc. Indirect AI risk policies would either affect AI but not explicitly govern it or address consequences of the use of advanced AI systems. This could include both

policies that affect AI and those that are AI-agnostic. For example, a policy that puts in place stronger protections for privacy in general would reduce the amount of training data available, and thus the speed of AI development, and could be considered an indirect approach. An AI-agnostic policy, for example, would be basic minimum income to address technological unemployment, which could be considered a risk if it leads to societal destabilization. AI risk relevant policies would affect neither AI nor the consequences of it but would rather make it easier for sound AI risk policies to be developed and implemented, such as changing the rules and procedures of government itself to alleviate the pacing problem.

There is another layer of classification that should be applied to AI risk policy based on Lowi's Typology [29]. Lowi categorizes policies into regulatory, distributive, redistributive, and constituency categories. Regulatory policies regulate one's behavior, restricting or incentivizing certain actions, such as the mandating of seat belts in cars. Distributive policies are policies that take money from the general treasury and use them for a specific project that directly benefits one group, such as a dam or research grants. Redistributive policies are those which fundamentally alter the distribution of wealth and resources in the whole of society, such as tax and welfare policies. Constituency policies are those that alter the composition and the rules and regulations of government, such as creating a new executive agency.

Each one of these typologies has with it a certain set of political conditions, as they impact people, businesses, and members of government differently. For example, both basic minimum income and the creation of AI safety standards are policies that are intended to reduce existential risk. However, both of these policies will have a different set of political pressures. Basic minimum income is a redistributive policy, which would move substantial amounts of wealth between classes of society. This would mean that it would likely become a nationwide controversial issue with two opposing camps based largely on who benefits and who loses. By contrast, AI safety standards are a regulatory policy, and while there would be two groups opposed to each other on the issue (unless it comes in the form of voluntary self-regulation by the industry), the political factors around it would look different. Regulatory policies are not usually salient or popular to the general public, and thus, the political battle would be largely limited to regulators, experts, and the business class. This typology will help us to understand how the different policies will be treated in the policymaking process. In other words, policy creates politics. Further work on developing this might be useful for understanding the likelihood of policies being adopted and could shift strategies for which policies to pursue.

4. The Policymaking Cycle

4.1. Problem Identification, Agenda Setting, and Policy Formulation

The first few steps of the policymaking process: Problem identification, agenda setting, and policy formulation, are usually tied together [30], including in a so-called 'multiple streams framework'. The multiple streams framework attempts to explain how policies reach the agenda when policy entrepreneurs are able to couple the policy, politics, and problems streams to open up a policy window, the opportune time when all the conditions are right to get a policy on the agenda [31].

4.1.1. Problem Stream

There are many problems in society. However, the public does not seek government intervention for many of these problems. There are some basic requirements for an issue in society to become a policy problem, which is that it is something that the public finds to be intolerable, government can do something about, and is generally seen as a legitimate area for government to work on [30]. Policy problems can also arise when there are two or more identifiable groups who enter into conflict in a policy arena for resources or positions of power [32].

The first condition for an issue to be considered a policy problem is that it is something that the public or a group finds to be intolerable. Indicators such as statistics can help to identify a problem.

These can be used objectively, for understanding conditions in society, or politically, when they are used to justify a political position: for example, using gun violence statistics as an argument for gun control. What is considered an issue over time changes because of the evolution of society. Changes in values, distribution of resources, technology, etc. will change what issues are considered in society [30]. In AI governance, identifiers such as the rate of technological progress or the proliferation of autonomous weapons could be used as examples. Creating a list of politically salient identifiers or metrics could be potentially useful for creating long-term strategies and goals.

How the issue is framed is very important for whether or not it will be considered a policy problem [30]. Is mandating seatbelts in cars beneficial for public safety? Or is it paternalistic? Are these problems legitimate for government to handle? The framing of a problem can have an overwhelming impact on whether or not it is considered a problem appropriate for government to even formulate policy on. It can also impact the content of the policy. Whether you define access to transportation for handicapped people as a transportation problem or a civil rights issue determines whether the acceptable solution involves buying special needs vans, or costly upgrades to buses and subways to ensure equal access. Framing can also raise the priority of a policy problem by, for example, calling it a crisis and raising a sense of urgency.

The question of framing is also incredibly important for AI governance. For example, would autonomous weapons make war more humane by removing humans? Or will it distance ourselves from the violence and make us more willing to use them? The AI governance community needs to think about how these issues ought to be framed, and the consequences of doing so.

In order for an issue to be a part of the system agenda, or what the public or specific communities are discussing, there must be a focusing event. Focusing events are specific events that draw attention to a problem in society and the reasons behind it. The Sandy Hook school shooting, for example, is a focusing event that drew attention to America's gun laws. Moreover, events that occur outside of sector-specific focusing events [31], or past policies on these issues, can have a large impact, especially on the types of solutions used. For AI governance, "Sputnik moments" such as AlphaGo beating Lee Sedol would be an example that drew considerable media attention and generated much discussion about the future of AI, especially in China [33].

Understanding how to exploit these events for the AI governance agenda will be key to generating support and getting policies on the agenda. It is also important to stay on top of these events to understand the direction society is heading in—and to pre-empt or avert less productive or dangerous framings that might feed into arms races [31]. For example, Yampolskiy (2018) details a list of past failures by AI-enabled products [34]. How could work like this be used to influence the problem-setting? Could other AI risk researchers expand on it and build that work into a more thorough project to be used to draw attention to AI risk? Or, could attempts such as this backfire and cause pre-emptive stigmatization or ineffective policies?

4.1.2. Politics Stream

The politics stream is the combined factors of the national mood or public opinion, campaign groups, and administrative/legislative change. Decision-makers in government keep tabs on the swaying opinions of the masses and interest groups and act in a way that promotes themselves favorably, changing items on the agenda to stay relevant and popular, and to obscure unpopular policy stances. Changes in administration, especially when there is a major shift in the ideological composition of the institution, have a strong impact on what is included or not included on the agenda [31].

In AI governance, and for people involved in advocating and implementing policies, maintaining a key eye on domestic and international politics will be key. Knowing when and what kind of policy to advocate for, and to whom, is crucial not only to saving time and energy, but also for legitimacy. Trying to sell a nationalistic administration on greater UN involvement will probably not help someone with furthering their policy proposals and may even damage their (and their coalition's) political

capital and cause. However, other forms of cooperation, such as bilateral cooperation for reducing the risk of accidents [35], may be more promising.

AI governance researchers will need to consider how the political landscape should shape their recommendations or policy proposals. Not only would it determine if their recommendations would ever get considered, but if it was implemented, how would it affect the national mood? Would the next administration simply walk it back? How would other interest groups react and impact the long-term ability to reduce risk? If administration changes result in a flip-flop of ideology, what does that mean for AI risk policies associated with the past administration? Could an AI risk policy group maintain influence throughout changing administrations? All of these have implications on our ability to reduce AI risk, and this means that the policymaking strategy will not only have to be robust but also flexible enough to survive changing political conditions.

4.1.3. Policy Stream

The policy stream, which is in essence the policy formulation aspect of the policy cycle, is the "soup" of ideas that are generated by policymakers [35] when deciding what to do about a problem. Different policy networks create policies differently, with different levels of innovativeness and speed [35]. Understanding these differences and examining their implications for the AI governance field might be useful to understand its long-term impact and the specific strategic routes it should take. In other words, how should the AI governance research field itself be organized in a way that promotes useful and relevant solutions?

Despite the staggering number of policy proposals coming out, only a handful will ever be accepted. These policies compete with one another and are selected on a set of criteria, which include technical feasibility, value compatibility [35], budgetary and political costs, and public acceptance. Policies that work will also be technically sound, with no major loopholes, and a clear rationale for how its provisions would lead to actually achieving the policy objectives [30]. This actually creates some key considerations for the field. It means that many ideas are either functionally useless due to their political limitations, unlikely to be adopted in the face of easier or less politically costly options, do not have viable policy mechanisms to achieve their goal, or are otherwise intractable prospects for government. Even if all of the above conditions are resolved, loopholes and unintended consequences may neuter the policy or make conditions worse. This vastly reduces the space of possible solutions. Further, even though the ability for policy implementation or values might change over time, it is still a matter of how much and when. This begs the question: What problems can be solved when, how, and by whom? What does that mean for the large picture strategic approach?

Where should our policies originate from? While there are a bunch of policy ideas out there, only a few are ever seriously considered for adoption. Sources of these policies include (in the United States Federal Government, for example) the President along with the Executive Office of the President, Congressional leaders, government agencies (mostly small incremental changes and adjustments), temporary organizations or 'adhocracies' that serve to investigate specific topics, and interest groups whose topical expertise and political power can sometimes make them de facto policymakers. Each of these areas have differing levels of legitimacy, influence, and degree to which they can make policy changes. A question to consider is not only where in the policy network AI risk policymakers should focus on making these policies, but where they can best advocate for the creation of additional bodies like adhocracies to create additional policies, and what implications that has for the field at large.

With regard to the policy formulation phase of policymaking, a continuum of political environments has been created such that on one extreme, there are policies with publics and on the other, there are policies without publics [36]. When policies are formulated, it is important to consider political environments relevant to the issue. The term "publics" refers to groups who have more than a passing interest in an issue or are actively involved in it. It appears that AI risks are issues where there are limited incentives for publics to form because of problems being remote, costly, or even abstract and uncertain. What does this mean for the AI safety community? How can interest groups be

created most effectively? How can these issues be best expressed so that they do not seem so remote, abstract, or uncertain?

4.1.4. Policy Windows and Policy Entrepreneurs

This framework assumes that policy decision-makers, the legislators and bureaucrats in government exist in a state of ambiguity, where they do not have a clear set of preferences, and each set of circumstances can be seen in more than one way. This cannot be resolved with more information, as it is not an issue of ignorance. The example that Zahariadis (2007) gives is that "more information can tell us how AIDS is spread, but it still will not tell us whether AIDS is a health, educational, political, or moral issue [31]".

Overall, the multiple streams framework describes government organizations as "organized anarchies" where institutional problems run rampant, there are often unclear or underdefined goals, overlapping jurisdictions, and a host of other problems that mean that decision-makers have to ration their time between problems and do not have enough time to create a clear set of preferences, make good use of information, or take the time to comprehend the problem for sound decisions on policies. In essence, decision-makers are not rational decision-makers by any stretch. Instead, it depends on the ability of policy entrepreneurs to couple the three streams and manipulate the decision-maker into achieving their intended policy goals [31].

Policy entrepreneurs, who are the policymakers, advocates, interest groups, etc. who push to make specific legislative changes in their areas, only have a short window of time to have their proposals added to the formal agenda. It is when the right political environment, a timely problem, and a potentially acceptable solution all meet together with a policy entrepreneur who can manipulate the situation to their advantage. Because decision-makers exist in a state of ambiguity, policy entrepreneurs are able to manipulate their interpretation of their information to provide meaning, identity, and clarity.

Policy entrepreneurs use different tools and tactics to manipulate the way decision-makers process information and exploit their behavioral biases. Framing tactics, for example, can be used to present a policy option as a loss to the status quo, not taking note of the degree of loss it creates, exploiting decision-makers who are loss-averse, and may push them towards more extreme options like going to war to make up for those small losses [31].

The manipulation of emotions through symbols and the identity or social status of a decision-maker can also pressure them to make certain choices; policies around flag-burning are a great example of this. Because decision-makers are under a great deal of stress and are time-constrained, the strategic ordering of decisions, or 'salami tactics', creates agreement in steps by reducing the total perceived risk of a policy [31]. The manipulation of symbols in the way that artificial intelligence is being framed today has already occured. At first, anti-autonomous weapons advocates were describing 'armed quadcopters' as a serious problem with little media attention [37]. These were rebranded as 'slaughterbots' and a short-film was released with substantial media attention. However, what sort of long-run impact will this have on the field? While giving policymakers straight facts and solutions seems appealing, AI risk policymakers have to consider that it is impractical in reality and may have to accept the inevitability, to policy success, of tactics like framing. Which begs the question, which tactics should they use and how? Questions like these must be considered.

All of this strongly requires an appropriate consideration. Consider, if there are some problems that can only be resolved through state action (such as an arms race), that means that it is dependent on the policymaking process, and thus, these solutions can only be passed when policy windows open. Therefore, how many of these opportunities do AI risk policymakers get? Or, how many chances do they get to implement AI risk policies? These windows only open every once in a while, and they are often in fragile conditions. For example, Bill Clinton's campaign in 1992 aimed to reform the healthcare system and made it a campaign priority, but his administration's failure to pass the bill closed the window [31]. In other words, what impact does this have on AI governance and policy implementation timelines and what does that mean for the field as a whole?

However, in order for a policy entrepreneur to manipulate decision-makers, they must have access to them, which is highly dependent on both the legitimacy of their issue but also for the legitimacy of the group itself and their interest. One of the ways that policy entrepreneurs increase their own influence is to create new decision-points that they can exploit and to reduce access of other groups [32]. AI risk policymakers and advocates will have to find some way to gain access to decision-makers. For example, working on near-term or non-existential risk issues with AI might help someone to build the social capital and network that is necessary to work on existential risks issues. This would not only make it easier people in the field to implement their solutions but to also make themselves gatekeepers to the decision-makers, which could help with preventing policies that would increase existential risks (whether from AI or other sources) from getting through. This may be an area that needs further research. Aspects such as a group's access to decision-makers, the advocating group's legitimacy, biases of the institution [38], and a group's ability to mobilize resources will determine what gets added to the agenda, and the AI risk community will need to work on building all of these. AI policymakers will need to develop a strategy for how to get the right people into the right places and how to coordinate between different groups.

Getting on the formal agenda is a competitive process because there are fundamental limits to a decision-maker's time, and because the policy may be perceived to harm the interests of other groups. Opposing groups can use a variety of tactics, such as denying that the problem exists, arguing that it is not a problem for government, or arguing that the solution would have bad societal consequences, to deny it agenda status. Other factors that could deny an issue agenda status include changing societal norms, political changes, or political leaders avoiding having to be confronted by an issue that hurts their interests. Thus, AI policymakers will need to know how to overcome and adapt to these changing situations and other organizations preventing their policies from being adopted.

AI governance and policy experts will need to pay attention to the arguments being used for and against superintelligence, and whether or not this will become a political issue. Baum (2018) notes that superintelligence is particularly vulnerable to what is known as politicized skepticism, skepticism that is not based on an intellectual disagreement about the problem, based on good-faith attempts to understand the arguments, but rather to shut down concerns based out of self-interest (or a conflict of interests). Some major AI companies, and even other academics, have criticized the idea of superintelligence out of what seems to be their own self-interest as opposed to genuine concerns [39]. This would have a devastating impact on AI policy advocates in a similar way that the tobacco industry significantly impacted scientific efforts to study the public health links between tobacco and cancer.

4.2. Policy Adoption

The next stage of the policy cycle is policy adoption, or when decision-makers choose an option that adopts, modifies, or abandons a policy. This does not necessarily take the form of choosing from a buffet of completed pieces of policy, but rather to take further action on a policy alternative that is more preferable and that is more likely to win approval. At this point, after much bargaining and discussion, the policy choice will only be a formality, or there will be continuous discussion and disagreement until there is a formal vote or decision made. This is an important field to analyze for AI policymakers for the obvious implication that they will want their policy proposals being chosen, and so they will need to understand and design strategies to do so. Further, as will be discussed later, when changes do occur, they can often bring with them wider changes in public policy [40], an implication that will need to be taken into account.

The advocacy coalition framework is a theory on policy adoption but also incorporates every other aspect of the policy cycle with it. The theory describes the interactions of two or more 'advocacy coalitions'; groups of people from a multitude of positions who coordinate together to advocate for some belief, or to implement some policy change (potentially over many fields) over an extended period of time [41]. These do not need to be a single, explicitly delineated organizations like the National Rifle Association but could include loosely affiliated groups of organizations and/or individuals, all

working towards the same goal. Building and maintaining coalitions will be one of the major tasks that AI policymakers will need to work on, and so, examining this framework will be highly valuable.

What is it that binds a coalition together? All advocacy coalitions share some form of beliefs. However, the advocacy coalition framework uses a hierarchical belief system. The deepest and broadest of these are deep core beliefs, which are normative positions on human nature, hierarchy of value preferences (i.e., should we value liberty over equality?), the role of government, etc. Policy core beliefs are the next stage of the hierarchy, which involves the extension of deep core beliefs into policy areas. Both of these areas are very difficult to change, as they involve fundamental values. This actually creates an issue where, due to differing fundamental and personal values which lead to lack of interaction, different coalitions often see the same information differently, leading to distrust. Each may come to see the other side as "evil", reducing the possibilities of cooperation and compromise [41].

The deeply held convictions of what a policy subsystem ought to look like are called policy core policy preferences and are the source of conflict between advocacy coalitions. They are the salient problems that have been the long-running issues in that area for a time. Policy core policy preferences shape the political landscape, dictating who allies with whom and who the enemies are, and what strategies coalitions take.

The final level of the belief hierarchy are secondary beliefs, belief that cover procedures, rules, and things of this nature. These are very narrow in scope and the easier to change, requiring less evidence and little bargaining to change.

Understanding the values and beliefs of different existing coalitions, groups, and individuals is key to building and maintaining new coalitions for AI policymakers. This brings up a few considerations. Since it is difficult for conflicting coalitions to work together, will AI policymakers have to choose certain coalitions to work with? What are the costs, benefits, and the potential blowback of this? Since some policies related to AI risk are not in a mature policy field (and thus do not have established coalitions), what can be done to shape the field beforehand to their advantage and/or promote cooperation among coalitions that are likely to form? Further, since secondary beliefs are relatively easy to change, what can be changed to help reduce existential risk?

On a macro-level, this AC Framework acts as a cycle. Relatively stable parameters, as mentioned before, exist in the status quo since policy arenas usually come to some equilibrium where one coalition dominates the policy subsystem. Then, policy changes made by an advocacy coalition or an outside event create a fundamental change in the world, whether it is a change in public opinion or in the rules and procedures governing a subsystem, which changes the initial stable parameters, such as a major event like a mass shooting. These lead to a shift in power that allows another coalition to gain influence over the types of policies being adopted. However, especially in the case of controversial legislature, policies that require multiple veto points to pass will create access for multiple coalitions. This means that even a coalition that dominates a subsystem will not have unilateral ability to dictate policies in some situations. Others, however, especially where there are few decision-makers or an exceptionally influential decision-maker, can result in highly monopolized systems. Questions such as how to be resilient to these changes in conditions, how to facilitate changes into conditions that are beneficial to AI policymakers, and how to construct policy subsystems in a way that is conducive to AI policymakers' goals are useful questions to consider.

This theory describes policy adoption on a very broad level, but how do the decision-makers themselves decide which policies to move forward with? Different incentives and restrictions come to play at different levels of policymaking. For example, highly salient and popular issues are more likely to be influenced by popular opinion, whereas obscure technical issues will likely be determined by policy experts in that field. Different factors that affect both individual and group decision-makers also come into play, such as their personal, professional, organizational, and ideological values. For legislators, their political party and their constituency also play an overwhelming role in their decision-making. Understanding and mapping out these factors will be necessary for the successful implementation of AI risk policy.

On top of these factors, decision-makers usually never have the time, expertise, or even care enough to be able to come up with a fully rational approach to deciding most policies. In many cases, legislators will seek out the advice of other legislators and experts and follow their lead. Due to this being a widespread practice, a few key institutions and leaders often have disproportionate power. For those working in AI risk policy, it is necessary to understand these things so that the message they craft for as to why policy change should occur, and whom to specifically target to get widespread adoption from other decision-makers in the policy arena.

4.3. Policy Implementation

Policy implementation is a key step in the policymaking process. It is defined as "whatever is done to carry a law into effect, to apply it to the target population … and to achieve its goals" [30]. In other words, it is the activity where adopted policies are carried into effect [30]. However, that is not to say that it is a very distinct step that can be clearly distinguished from others. Every implementation action can influence policy problems, resources, and objectives as the process evolves [42]. Policy implementation can influence problem identification, policy adoption, etc.

Two broad factors that have been offered for the success of policy are local capacity and will [42]. In other words, is there enough training, money, and human resources, along with the right attitudes, motivation, and beliefs to make something happen? It is suggested that the former can be influenced much more easily than the latter as more money can be received and consultants can be hired. For AI risk, both questions are relevant: How to increase capacity and how to influence the influencers. With the former, it has been estimated that about $9-$20 million is currently spent on AI risk [43,44]. With the latter, studying the opinion of the public as well as experts might be a useful approach. One survey [45] indicates that only 8% of top-cited authors in AI consider that human-level AI would be extremely bad (existential risk) for humanity. Another survey that is more recent [46] indicates that machine learning researchers think on average (median) that there is a 10% probability that human-level machine intelligence will result in a negative outcome and 5% probability that it will have an extremely bad outcome (existential risk). The general public seems to be generally cautious, with a survey showing 82% of Americans believing that AI/robots should be managed carefully [47].

This part of the policymaking process is very difficult as the literature is generally quite pessimistic about the ability of policies to bring social changes into effect [48]. However, the authors of the cited paper have identified conditions of effective implementation based on successful examples. These conditions are (a) the policy is based on a sound theory of getting the target group to behave in a desired way, (b) policy directives and structures for the target group are unambiguous, (c) the leaders implementing the policies are skillful with regard to management and politics and committed to the goals, (d) policy is supported by organized constituency groups and key legislators as well as courts throughout the implementation process, and (e) the relative priority of policies is not significantly undermined over time by other policies or socioeconomic changes. Additionally [49], having carefully drafted statute that incentivizes behavior changes, provides adequate funds, expresses clearly ranked goals, is an implementation process, and has few veto points is also vital to the success of a policy.

With regard to AI governance, the ambiguity and complexity of the problem creates a major hurdle for effective policies to be developed. These problems are nonlinear, very hard to predict, and may have the traits of wicked problems in the sense that solving one problem can create new problems. Breaking down AI risk policy into multiple domains as discussed in the previous section helps with creating somewhat less ambiguous objectives, such as changing the education system to be more conducive for technological growth. Even then, however, because many of the issues are either complex or have not happened yet, it is difficult to create concrete objectives and policies. AI risk is not like noise pollution, where there is an easily identifiable, manageable, and tractable problem. Further research could help to identify concrete and tractable issues that might lead to a reduction of risk. In addition, when trying to develop and implement policy, AI policymakers will need to keep in mind

factors such as to what extent there is support for it in the executive branch, with outside organizations, and how exactly the policy is written and how those change throughout the policymaking cycle.

Another key consideration for successful policy implementation that was identified from the literature is engaging with the community to increase readiness to accept and devote resources to policy-related problems. It has been acknowledged that there are no good evidence-based ways of achieving community buy-in. This is an area that might be useful to study in order to increase the chances of successful reduction of AI risk. There are different stages of community readiness, such as no awareness, denial, and vague awareness to preplanning, preparation, initiation, and stabilization phases [49]. It is important to understand what counts as the community and what phases different subcommunities of AI safety field are in. Earlier, this paper mentioned a survey about AI experts and showed that their readiness with AI risks was low. Other relevant experts, the public, and other types of subcommunities might have different levels of readiness.

It has been suggested that "the more clearly the core components of an intervention program or practice are known and defined, the more readily the program or practice can be implemented successfully" [49]. In other words, policies and steps of implementation of those policies have to be very clearly expressed. What implications does this have for AI risk? Researchers and policymakers should evaluate how clearly core components have been expressed in this field and improve them as necessary.

4.4. Policy Evaluation

The final step in the policymaking cycle is policy evaluation. This includes activities related to determining the impact of the policy, whether it is achieving its goals, whether the rules and procedures it lays out are being followed, and other externalities or unintended consequences [30]. As we have explained before, policy evaluation does not have to occur only at this step. For example, the impact of a policy is estimated already in the early stages. Anderson et al. highlighted different types of policy evaluations in their book but especially considered systematic evaluations of programs. This involves "the specification of goals or objectives; the collection of information and data on program inputs, outputs, and consequences; and their rigorous analysis, preferably through the use of quantitative or statistical techniques" [30]'.

Policy evaluation examines a policy to understand its impacts in multiple ways [30]. First, is the policy affecting the population that it is intending to target? In AI risk policy, this could be anything from large tech companies, to AI researchers, to people affected by technological unemployment. Second, are there populations that are being affected that were not intended? These externalities could be positive or negative. Third, what are the benefits and costs associated with this policy? AI policymakers will want to ensure that their policies actually reduce risk and that the costs are not so astronomical that they become politically infeasible. Finally, what long-term costs and benefits does a policy have? This is especially important for AI risk policy, as decisions now could have a major impact on the long-term risk that AI has. In AI governance and policymaking, research needs to be done on what sort of indicators or metrics are used for the reduction of risk, and for identifying what goals that should be achieved.

If the previous steps in the policymaking process have generated goals that are unclear or diverse, it is very difficult to evaluate the impact of the policy [30]. Different decision-makers can more easily reach a differing conclusion about the results of a program in that case, or may not follow it all [30]. How the goals of an AI risk program are defined is, therefore, very important.

Another key consideration for policy evaluation is how to make sure that the results are objectively measured. Agency and program officials may be wary of possible political consequences of the evaluation process [30]. If it turns out that the program was not useful or even detrimental, this might have consequences to their influence and career. Because of this consideration, they might not be very interested in correct evaluation studies or they may hinder the process in some other way. There are many ways an evaluation of a policy might be ignored or attacked, such as claiming it was poorly

done, the data were inadequate, or the findings inconclusive [30]. Thus, it is important that researchers are provided with high-quality and relevant data-sets that are accurate.

There is also the distinction between policy outputs and outcomes [30] to consider. Outputs are tangible actions taken or things produced, such as collecting taxes or building a dam. Outcomes, on the other hand, are the consequences for society, such as lower disposable income or cleaner air quality. Outputs do not always produce the intended outcomes, which is highly evident in areas such as social welfare policy, where policies may unintentionally trap people in poverty. For AI policymakers, it is very important to consider whether their policy outputs will have the intended consequences, and if so, how to correct that policy.

The evaluation of a policy and the political responses to it can result in the termination of it [30]. Assuming that AI risk policymakers do not want their policies to be terminated or altered in a detrimental way, how can they make sure this does not happen? A policy getting altered to be more effective might be a good thing, but termination can bring unpleasant and negative connotations. It might even have negative consequences to the community [30]. What exact consequences might it have politically? Further, it is important to remember that many policymakers' time horizon only goes until the next election, and so, they often seek immediate results, often before the returns come into fruition. While this may not impact all policies, as this mostly applies to salient policies like healthcare and education, AI policymakers should keep this in mind and try to understand how it might impact their work.

5. Conclusions

There are multiple policy options that could be chosen that either directly or indirectly reduce AI risk, or relevant policies that could help with further efforts to reduce AI risk. Because different policy arenas have different political conditions, and the policymaking process itself draws a number of important challenges, this brings up questions as to what policies in what order are chosen, what strategies are used to get these policies passed and implemented by the government, and the larger impact of these choices on AI governance and risk as a whole. This paper argues that a new subfield of AI governance research on AI policymaking strategies should be further investigated to draw implications for how these policies should be designed, advocated for, and how organizations should approach solving this issue.

6. Limitations and Future Research

This paper is intended to be a broad overview and to be a conversation starter for future research into this area. Thus, there is a strong limitation to the depth of research in this paper. However, it is expected that future work will be done to further refine the line of thinking laid out above, along with further in-depth study into the different theories and their applicability to AI risk.

One of the major limitations of this paper is that the stages heuristic presented in this paper has been heavily criticized and is subject to debate about its effectiveness. Sabatier (2007) has criticized it for not being a causal theory, having a strong top-down bias, among other critiques. However, he also notes that there is much up to debate, with some scholars such as Anderson (2010) advocating for it. There are also a number of other theories that were not discussed in this paper, such as Institutional Rational Choice, the punctuated equilibrium framework, the policy diffusion framework, and other lesser-known theories. Future research is expected that will explore which policy frameworks should be focused on in AI risk research.

The other limitation of this paper is that its applicability to the international governance of AI was not discussed. Future research that looks at how much these theories apply to foreign policy and the international governance of AI in general would be useful. If these theories have a very limited or no impact on the international governance of AI, then figuring out how much work can be done to reduce AI risk in domestic policy would determine the usefulness of these theories.

Throughout the paper, a number of key considerations have been raised. For convenience, a list of them has been curated below below.

7. Summary

This part of the paper summarizes and lists some of the key questions and considerations brought up in the discussion.

Thesis level consideration:

- How do the politics and administrative mechanisms of policymaking affect how policies to mitigate AI risk are created and implemented?

Considerations from Typologies of Policies:

- Are there AI risk policies that should be implemented first? What are the methods to decide this?
- What types of policies should the AI risk policymakers try to get implemented? Why should those types be prioritized?
- What are the political considerations surrounding different sets of policies, and how does that affect their ability to be implemented?

Considerations from Problem Identification, Agenda Setting, and Policy Formulation:

- Is this issue or policy legitimate?
- Would the policy be supported by the current administration and be able to be maintained through changing administrations?
- Which policies out of different sets of potential solutions are politically feasible?
- Are there less costly alternative policies that AI risk policymakers will have to compete with?
- How does attention to problems by different communities affect AI risk policymakers' actions?
- What types of framing of policy issues are most beneficial? What types are most dangerous?
- Is there a way to determine how framing will determine policy content?
- What focusing events have occurred in the field of AI?
- How can AI risk policymakers utilize focusing events to further policy agendas?
- What effect do other organizations have on reducing the legitimacy of AI risk?
- What can be done to respond to these counter-movements effectively? What kind of responses to objections are most convincing?
- How many policy windows will there be for a particular issue? What does this mean for AI risk policymakers' overall strategy?
- What role should AI risk policy entrepreneurs play in AI governance?
- How and where should AI risk policy entrepreneurs gain access in government?

Considerations from Policy Adoption:

- What policy alternatives are more likely to win approval to improve the odds of success for AI risk reduction?
- What strategies can be used to improve the chances of a preferred policy to be adopted?
- Which groups or individuals could join AI risk coalitions, what criteria are used to decide this, and what costs does them joining the coalition have?
- What role can organizations outside of AI risk play in furthering AI risk policymakers' agenda?

Considerations from Policy Implementation:

- Is this solution technically feasible for governments to implement?

- Are there enough resources, will, and support by leaders and constituency groups to be successful in implementation?
- Is the policy crafted in a way that effectively structures incentives for the target group?
- Is the policy unambiguous? If so, then how will that affect its ability to be implemented?
- Are the goals of the policy in conflict with any other policy or changes in society?
- Are there any veto points in the policy's statutes to prevent effective implementation?
- How will the contents or the political factors surrounding of a policy be affected during implementation?
- Do the relevant communities accept the issue, and are they willing to devote resources to resolve it?

Considerations from Policy Evaluation:

- Are the policy outputs having the intended outcomes?
- What are the consequences of any unintentional outcomes?
- What are the political factors surrounding the metrics that are being used to evaluate the policy?
- Do the political costs or benefits of the policy have an impact on its success?
- If the policy is terminated, will there be any negative political consequences?
- How can AI risk policymakers update the policy? How can they prevent changes by other groups that would be harmful?
- How will the limited time horizons of lawmakers and other groups affect the evaluation of the policy?

Author Contributions: Conceptualization, B.P. and R.U.; methodology, B.P. and R.U.; writing—original draft preparation, B.P. and R.U.; writing—review and editing, B.P. and R.U.; project administration, B.P. and R.U.

Funding: This research received no external funding.

Acknowledgments: The authors thank Matthijs Maas, Seth Baum, Sabrina Kavanaugh, Max Daniel, and the organizers and participants of the AI Safety Camp for useful comments and feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References and Notes

1. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
2. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK; New York, NY, USA, 2014.
3. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute: Oxford, UK, 2018. Available online: https://www.fhi.ox.ac.uk/govaiagenda/ (accessed on 17 December 2018).
4. Baum, S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. 2017. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741 (accessed on 11 November 2019).
5. Everitt, T.; Lea, G.; Hutter, M. AGI Safety Literature Review. *arXiv* **2018**, arXiv:1805.01109.
6. Joy, B. Why the future doesn't need us. *Wired* **2000**, *8*, 238–263. Available online: https://www.wired.com/2000/04/joy-2/ (accessed on 6 January 2019).
7. Hibbard, B. *Super-Intelligent Machines*; Springer: New York, NY, USA, 2002.
8. Hughes, J.J. Global technology regulation and potentially apocalyptic technological threats. In *Nanoethics: The Ethical and Social Implications of Nanotechnology*; Allhoff, F., Ed.; John Wiley: Hoboken, NJ, USA, 2007; pp. 201–214.
9. McGinnis, J.O. Accelerating AI. *Northwest. Univ. Law Rev.* **2010**, *104*, 366–381. Available online: https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=nulr_online (accessed on 14 March 2019). [CrossRef]
10. Scherer, M.U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. J. Law Technol.* **2016**, *29*, 354–398. [CrossRef]
11. Guihot, M.; Matthew, A.F.; Suzor, N.P. Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt J. Entertain. Technol. Law* **2017**, *20*, 385–456.

12. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
13. Yampolskiy, R.; Fox, J. Safety Engineering for Artificial General Intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
14. Erdelyi, O.J.; Goldsmith, J. Regulating Artificial Intelligence: Proposal for a Global Solution. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), New Orleans, LO, USA, 2–3 February 2018. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3263992 (accessed on 6 January 2019).
15. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Environ. Law J.* **2013**, *31*, 307–364.
16. Shulman, C. Arms control and intelligence explosions. In Proceedings of the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, 2–4 July 2009.
17. Goertzel, B. The Corporatization of AI is a Major Threat to Humanity. h+ Magazine. 2017. Available online: http://hplusmagazine.com/2017/07/21/corporatization-ai-major-threat-humanity/ (accessed on 6 January 2019).
18. Bostrom, N. Strategic Implications of Openness in AI Development. *Glob. Policy* **2017**, *8*, 135–148. [CrossRef]
19. Dewey, D. Long-term strategies for ending existential risk from fast takeoff. In *Risks of Artificial Intelligence*; Müller V.C., Ed.; CRC: Boca Raton, FL, USA, 2015; pp. 243–266.
20. Goertzel, B. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? *J. Conscious. Stud.* **2012**, *19*, 96.
21. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Available online: https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/1c6q2kc4v_50335.pdf (accessed on 6 January 2018).
22. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Cave, S. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA, 27–28 January 2019.
23. Calo, R. Artificial Intelligence Policy: A Primer and Roadmap. 2017. Available online: https://ssrn.com/abstract=3015350 (accessed on 6 January 2019). It should also be noted that Calo is dismissive of the risk of artificial general intelligence.
24. Cave, S.; ÓhÉigeartaigh, S.S. An AI Race for Strategic Advantage: Rhetoric and Risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019. Available online: http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf (accessed on 14 March 2019).
25. Flynn, C. Personal Thoughts on Careers in AI Policy and Strategy. Effective Altruism Forum. 2017. Available online: https://forum.effectivealtruism.org/posts/RCvetzfDnBNFX7pLH/personal-thoughts-on-careers-in-ai-policy-and-strategy (accessed on 6 January 2019).
26. The specifics issues will depend on the type of government. For example, the types of difficulties would be different in a democracy vs. a dictatorship. This paper however will focus on federal republics.
27. Thank you to Sabrina Kavanagh for suggesting the idea that the policy process could inspire new ideas for AI governance researchers.
28. Brundage, M.; Bryson, J. Smart Policies for Artificial Intelligence. *arXiv* **2016**, arXiv:1608.08196.
29. Lowi, T.J. Four Systems of Policy, Politics, and Choice. *Public Adm. Rev.* **1972**, *32*, 298–310. [CrossRef]
30. Anderson, J.E. *Public Policymaking: An Introduction*, 7th ed.; Cengage Learning: Boston, MA, USA, 2010.
31. Zahariadis, N. The Multiple Streams Framework: Structure, Limitations, Prospects. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P., Eds.; Westview Press: Boulder, CO, USA, 2007.
32. Cobb, R.; Elder, C.D. What is an Issue? What Makes an Issue? In *Participation in American Politics: The Dynamics of Agenda Building*; Johns Hopkins University Press: Baltimore, MD, USA, 1983; pp. 82–93.
33. Allen, G. China's Artificial Intelligence Strategy Poses a Credible Threat to U.S. Tech Leadership. Center for Foreign Affairs Blog. Available online: https://www.cfr.org/blog/chinas-artificial-intelligence-strategy-poses-credible-threat-us-tech-leadership (accessed on 26 February 2019).
34. Yampolskiy, R. Current State of Knowledge on Failures of AI Enabled Products. Report. Consortium for Safer AI. 2018. Available online: https://docs.wixstatic.com/ugd/ace275_0ea60fe9b665439bb0b37d20beb89b6f.pdf (accessed on 6 January 2018).

35. Danzig, R. *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*; Center for New American Security: Washington, DC, USA, 2018. Available online: https://www.cnas.org/publications/reports/technology-roulette (accessed on 24 March 2019).
36. May, P.J. Reconsidering Policy Design: Policies and Publics. *J. Public Policy* **1991**, *11*, 187–206. [CrossRef]
37. Russell, S.; Aguirre, A.; Conn, A.; Tegmark, M. Why You Should Fear "Slaughterbots"—A Response. IEEE Spectrum. 2018. Available online: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-you-should-fear-slaughterbots-a-response (accessed on 9 January 2019).
38. Yudkowsky, E. Cognitive Biases Potentially Affecting Judgment of Global Risks. In *Global Catastrophic Risks*; Bostrom, N., Ćirković, M.M., Eds.; Oxford University Press: New York, NY, USA, 2008; pp. 91–119.
39. Baum, S.D. Superintelligence Skepticism as a Political Tool. *Information* **2018**, *9*, 209. [CrossRef]
40. James, T.L.; Jones B.D.; Baumgartner, F.R. Punctuated-Equilibrium Theory: Explaining Stability and Change in Public Policymaking. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P.A., Ed.; Westview Press: Boulder, CO, USA, 2007; Chapter 6.
41. Sabatier, P.; Weiblle, C.M. An Advocacy Coalition Framework. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P.A., Ed.; Westview Press: Boulder, CO, USA, 2007; Chapter 7.
42. McLaughlin, M.W. Learning From Experience: Lessons From Policy Implementation. *Educ. Eval. Policy Anal.* **1987**, *9*, 171–178. [CrossRef]
43. Farquhar, S. Changes in Funding in the AI Safety Field. 2017. Available online: https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field (accessed on 6 January 2019).
44. MacAskill, W. What Are the Most Important Moral Problems Of Our Time? TED Talk. 2018. Available online: https://www.ted.com/talks/will_macaskill_how_can_we_do_the_most_good_for_the_world (accessed on 6 January 2019).
45. Müller, V; Bostrom, N. Future progress in artificial intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Synthese Library; Springer: Berlin, Germany, Forthcoming 2014. Available online: https://nickbostrom.com/papers/survey.pdf (accessed on 6 January 2019).
46. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv* **2017**, arXiv:1705.08807.
47. Zhang, B.; Dafoe, A. Artificial Intelligence: American Attitudes and Trends. January 2019. Available online: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf (accessed on 3 January 2019).
48. Sabatier, P.; Mazmanian, D. The Conditions of Effective Implementation: A Guide to Accomplishing Policy Objectives. *Policy Anal.* **1979**, *5*, 481–504. [PubMed]
49. Sabatier, P.; Mazmanian, D. The Implementation of Public Policy: A Framework of Analysis. *Policy Stud. J.* **1980**, *8*, 538–560. [CrossRef]