Towards Safe, Ethical and Beneficial Artificial Intelligence in the European Union and Beyond

Towards Safe, Ethical and Beneficial Artificial Intelligence in the European Union and Beyond: A Multifaceted Framework for Governance

by

Charlotte Stix

A thesis submitted for the degree of
Doctor of Philosophy
Spring 2023

Eindhoven University of Technology
Department of Industrial Engineering & Innovation Sciences

Towards Safe, Ethical and Beneficial Artificial Intelligence in the European Union and Beyond: A Multifaceted Framework for Governance

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op 20-04-2023 om 11 uur.

door

Charlotte Stix

geboren te Wenen, Oostenrijk

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	Prof. Dr. C. Snijders
1e promotor:	Prof. Dr. V. C. Müller (FAU Erlangen-Nürnberg)
copromotor(en):	Dr. E. E. O'Sullivan
leden:	Prof. Dr. A. Wynsberghe (Universität Bonn)
	Prof. Dr. H. Haukkala (Tampere University)
	Prof. Dr. L. M. M. Royakkers

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Table of Contents

Abstract

This thesis puts forward novel frameworks with regards to ethically informed AI governance for both academia and the policy community in the EU. In doing so: (1) it provides relevant context and investigates the international proliferation of the term 'trustworthy AI', as advanced by the EU in government discourse, suggesting that the EU currently holds a "first mover" advantage in this space; (2) it conducts an in-depth analysis and review of key policy, investment and regulatory decisions informed by ethical considerations in the EU and what this may mean for the future; (3) it advocates for the concept of 'Actionable Principles' and proposes a number of elements to develop a suitable mechanism to achieve them; (4) it advocates for the application of an 'incompletely theorized agreement' for the purpose of achieving a sufficiently cohesive and strong AI policy and scholarly community, regardless of diverging perspectives on AI impact and finally, (5) it advocates for the importance of institution building as one component of achieving future impact on AI governance and puts forward a blueprint for potential future organizations, many of which would be tasked with the implementation and execution of the aforementioned projects.

AI governance in the EU is a multifaceted and evolving field: it is in need of further research and perspectives across a multitude of areas. I aim to contribute to several of these perspectives and hope that future researchers expand on these areas, in cooperation with technical researchers, ethicists, lawyers and government officials.

Introduction

1. Why do we need to think about ethically informed AI governance?

Recent years have seen a steady increase in scholarly (Coeckelbergh, 2020; Maas, 2021; Müller, 2020a) and policy work (Askell et al., 2019; Buyers, 2018; Cihon, 2019; Coeckelbergh, 2020; Cremer & Whittlestone, 2021) that (a) delineates ethical, societal and technical concerns about AI and (b) proposes a variety of solutions to those concerns. Various approaches have been proposed as to how society can ensure that AI systems are developed and deployed in a manner that evades negative downstream effects while increasing positive impact (Brundage et al., 2020; Bryson, 2018; Fischer et al., 2021; Turner, 2018). In particular, there has been a surge in academic literature examining ethical consideration with regards to AI (Boddington, 2017; Fjeld et al., 2020; Floridi et al., 2018; Jobin et al., 2019), mostly in direct response to — or anticipation of — real-world events. The academic literature in the field has equally indicated a need for a clearer framework to be implemented within which AI technologies can be governed to avoid and minimize existing and future shortcomings (Buolamwini & Gebru, 2018; Cihon et al., 2021; Cowls et al., 2021). The ubiquity of AI across sectors, as well as recent advances in the capability of general-purpose AI systems[1] (Amodei et al., 2016; Gruetzemacher & Whittlestone, 2019; O'Keefe et al., 2020), underline the importance of developing an overarching model and methods to ensure the technologies' beneficial coexistence and alignment with the values of the society in which they are deployed (Ess, 2006; Gabriel, 2020). Establishing such a model is not a straightforward feat. Value alignment (Christian, 2021), the known and unknown range of incidents one must

[1] "General-purpose AI" has no clearly defined technical definition, so various operational definitions are used depending on the given context and the utility of operationalizing in one way or another. OpenAI, one of the leading companies developing these AI systems, defines general-purpose AI as "highly autonomous systems that outperform humans at most economically valuable work".

safeguard against,[2] and the cross-border nature of these AI systems renders this a complex task in dire need of investigation and meaningful action. As indicated, there exists a wealth of academic scholarship outlining the societal impacts and near-term effects of AI (Crawford, 2021) and providing an ethical analysis (Jiang et al., 2021; Ryan & Stahl, 2020). This thesis, in comparison, approaches these topics through the lens of AI governance. It pulls together existing research to inform a future-oriented framework for ethical AI governance, specific to democratic governments with a focus on the European Union (EU).

Although the discussion of AI ethics is not new (Samuel, 1960; Wiener, 1960), advances in technology have inspired a more recent surge in proposed policy measures (Müller, 2020b). AI systems have been able to automate (and in some cases improve upon) human performance in narrow but domain-transferable tasks such as pattern recognition, anomaly detection, prediction, optimization, and autonomous operation. While the breadth of its potential application underlies much of AI's potential for doing good (Livingston & Risse, 2019; Rolnick et al., 2019; Vinuesa et al., 2019), it also means that AI raises issues in almost every sphere of human activity and society. The space of ethical and policy concerns raised is vast (Müller, 2020b; Tasioulas, 2019). Important issues that fall within the overlap of technical, ethical and policy concern include: algorithmic bias (Barocas & Selbst, 2016; Berk et al., 2018; Crawford et al., 2019), transparency and explainability (Doran et al., 2017; Gebru et al., 2021), the use of social robotics, the safety of autonomous vehicles (Anderson et al., 2016; Nyholm & Smids, 2016), or the potential of AI systems to be used in (or susceptible to) malicious or criminal activities (T. King et al., 2018). Further challenges are anticipated in the near future, as societies may increasingly have to reckon with: (a) privacy concerns in the face of widespread surveillance (Calo, 2010; Gasser, 2016), (b) the economic and political effects of technological unemployment (Danaher, 2019; Frey & Osborne, 2017), (c) the erosion of the global legal order by the comparative empowerment of

[2] See: https://partnershiponai.org/workstream/ai-incidents-database/.

authoritarian states (Danzig, 2017; Deeks, 2020; Maas, 2019), and (d) the possibility that the military use of AI technology in war could create new legal or ethical problems, operational risks, or upset strategic balances of power (Garcia, 2018; Horowitz, 2018, 2019; Horowitz et al., 2019; Lewis et al., 2016).

Another key issue, especially in the context this thesis' focus on the EU, is the potential for AI to erode democracy by way of 'computational propaganda' or 'deepfakes' (Chesney & Citron, 2019; Helbing et al., 2017; Nemitz, 2018). For example, in the ongoing war in Ukraine, Russia launched a deepfake video of president Volodymyr Zelensky announcing his surrender.[3] Although this deepfake was quickly spotted due to its unnatural appearance, future deepfakes may well not be and could influence wars and high-stakes decisions alike.

Malicious use, or misuse, is an unfortunate byproduct of AI development and deployment, even if the developers' intentions (for the product) are 'good'. This creates tricky ethical, technical and policy scenarios and echoes a familiar adage: "with great power comes great responsibility." This is scarily well illustrated by another recent case of (well-intended) misuse. In 2022, a group of scientists purposefully flipped the reward function on an AI system used for drug discovery. Within six hours, the altered AI system proposed 40,000 molecules that appeared to include known or plausible chemical weapon agents, illustrating "how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons" (Urbina et al., 2022). The risk of misuse is especially salient with DeepMind allowing AlphaFold to be open sourced.[4] Both areas — deepfakes and whether or not to regulate research and development — are ongoing topics of discussion when it comes to regulating AI in the EU and have been discussed at length in scholarly discourse (Helbing et al., 2019; Nemitz, 2018; Whittlestone & Ovadya, 2019).

[3] See: https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.
[4] Certain capabilities of AlphaFold have been limited, possibly due to misuse potential. See: https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe.

All this is to say that a select number of issues regarding the development and deployment of AI systems which have long spurred ethical debate among scholars have now become concrete topics within policy making. Despite existing and unresolved issues, technological progress continues and governments are forced to address an increasing number of (at times, avoidable) failures involving AI systems, and therefore grapple with translating many of the ongoing discussions within AI ethics into government policy and strategy. It is this combination of (a) existing academic literature, (b) real ripple effects of AI incidents and (c) a dramatic shift in the capabilities of AI systems that has rightly caused ethically-informed AI governance to become a feature of the field, deserving of further strengthening.

AI governance concerns itself with the development and implementation of the framework (Calo, 2017; Gutierrez & Marchant, 2021), methods (K. G. Greene & Gretchen Greene, 2022; Walz & Firth-Butterfield, 2019) and measures (Jelinek et al., 2020; Schiff et al., 2020) various actors can use to steer the direction of AI development and deployment. For the purpose of this thesis, the concept of governance will be explored in relation to the actions governmental actors may take, ranging from drawing up agreements, proposing regulations, developing subject specific policy,[5] and undertaking standardization efforts to set up new institutional infrastructures. The field of AI governance is relatively nascent and many published works are just starting to grapple with applied governance questions (Bollock, 2022). Concurrently, governments are laying novel groundwork to tackle AI, often directly reliant on previous academic and ethical investigation, as described in Chapters II and III of this thesis. Overall, the timing and focus of this thesis matches the pace of scholarly work and policy developments, aiming to provide relevant and actionable insights and frameworks. This thesis draws from scholarly work (ranging from AI ethics to political science), governmental policy and

[5] This means that AI policy is considered as a subset of AI governance. It is one of the predominant levers governments can use to steer AI development and deployment, in particular.

political efforts. Without claiming any one of these areas as the main field of investigation, it instead attempts to draw them together, bridging those fields and putting forward a broadly applicable yet modular framework for ethically informed AI governance.

Extrapolating from the pace of development of AI systems, it is likely that AI will become a deeply transformative technology for humankind (Russell, 2019), impacting all ways of life as we know them. It is, therefore, important to set meaningful safeguards now, to ensure that path dependencies unfold to allow a sufficient degree of control, oversight and cooperation between a number of key actors, contributing to beneficial outcomes. So, one of the core questions motivating this thesis is: *how can governments and governance practitioners ensure meaningful safeguards for AI development and deployment*?

It is difficult, if not impossible, to know with a high degree of certainty what measures must be implemented to avoid catastrophic failures or unforeseen negative impacts of AI systems in the coming years and decades. This is why implementing good governance measures matters, now. Indeed, even simple algorithms can have devastating impacts on individuals (T. C. King et al., 2020; Stanley, 2019), marginalized groups (Buolamwini & Gebru, 2018; Raso et al., 2018) and democratic institutions (Nemitz, 2018). Countermeasures and research on these effects almost always happen post fact, demonstrating how challenging it is to predict an AI system's functionality and side effects (Brundage et al., 2018). The post fact nature of our countermeasures also highlights the fragile stability which we seemingly still take for granted as a society, namely that we are and will be in a position to mitigate those effects *after* they have happened and after we have become aware of them. The success of this manner of working relies on an assumption that all incidents are salvageable or can be successfully known a priori and avoided. Given the ubiquitous nature of AI systems and their increasing capabilities (Grace et al., 2018; Pennachin & Goertzel, 2007), it is highly unlikely

that this method of working will suffice in the not-so-distant future.

It should be reasonably clear, however, that we cannot know all possible negative impacts of AI systems before they happen and, consequently, that we cannot be certain what a comprehensive and complete framework to avoid negative outcomes would look like. This thesis grapples with the feasibility of establishing such a framework. By drawing on a variety of fields and approaches, it looks to propose a number of considerations and scaffolding that could contribute to a more complete, agile and future-proof AI governance framework. The chapters and associated investigations follow a layered approach that could be followed in order to think through a select number of considerations. Equally, they can function as independent 'blueprint structures' and proposals. This is intentional. While the goal is to address a specific academic and policy audience familiar with these issues, I hope that those who are just familiarizing themselves with these issues can equally benefit from these chapters. They are intended to be self-contained and suitable to inform interested parties (such as government officials) at many levels of expertise.

To recapitulate, the intent of this thesis is to provide an agile and coherent set of measures and considerations for policymakers, AI governance scholars, and other relevant practitioners to draw on to think about or develop an adaptable AI governance framework, informed by ethical considerations. For this thesis to have its intended impact, the measures, approaches and considerations proposed herein, while informed by academic research, need to be practical in their application and applicable in existing real-world settings. All chapters, therefore, examine their central concern through a practical lens. They examine, for example, a lack of operationalizability of AI ethics guidelines and principles, turf wars between communities working on AI ethics and policy concerns with diverging timelines and the calcified infrastructure through which many AI governance efforts will come into force. The chapters offer concrete steps to address these topics. They also elaborate on a range of considerations that AI governance practitioners could use to

address underlying cruxes.

The primary audience for this thesis are academics working on AI governance and policy practitioners. The latter includes, but isn't limited to, policymakers, regulators, government officials and those groups and individuals lobbying governments regarding the steering of AI progress and impact. Each chapter addresses a different subgroup. For example, Chapter III focuses on the interplay between policymakers and the broader AI community. Chapter IV focuses on academic scholars and their diverging projections of the future-impacting policy efforts and Chapter V focuses on governments broadly defined.

Another aspect of the thesis' academic contribution is motivated by the understanding that, in this particular subfield, much of the most important academic work that has been recently published, while often quite technical and subfield-specific, is relevant to a broad academic AI governance community, including policymakers and governments. From a purely academic perspective, this thesis contributes to: (1) the mapping and understanding of interdisciplinary AI governance research on international decision-making to increase societal benefice in Chapter I (Feijóo et al., 2020); (2) structural reviews of AI-ethics-adjacent policy efforts in Chapter II (Roberts et al., 2021); (3) a range of framing and mapping efforts for the intersection of AI ethics and guidelines in Chapter III (Madaio et al., 2020); (4) disentangling and reframing timeline concerns with regards to AI ethics and AI governance issues in Chapter IV (C. Prunkl & Whittlestone, 2020) and (5) to the study of institution-building applied to AI governance in Chapter V (Cihon et al., 2020). Moreover, the chapters of this thesis serve to strike connections (i) between different communities within AI ethics scholarship, (ii) between academic AI ethics scholarship and existing literatures on political science and institutional design and (3) between AI ethics scholarship and AI governance practitioners, which must be a mutual relationship if the intention is to ensure good governance for AI. This field is not just about translating philosophical insights into terms

policymakers can work with; it is also about making AI ethics scholars more aware and mindful of the practical constraints and opportunities that face AI governance initiatives in the world.

Despite differing opinions on AI progress (Müller & Bostrom, 2016) and thus the capabilities one should prepare for, there is a general agreement that we should avoid the development and deployment of AI that where the harm to individuals, sentient beings, society at large or the environment outweighs the benefits (see further discussion in Chapter I). However, the prescription of concrete actions to ensure that the technology (and the humans involved in developing and deploying it) adhere to this general sentiment has been an ongoing topic of discussion for ethics, policy and, most recently, legislation.

2. The difficulty of future-oriented governance interventions

Making decisions and policies that take the future and future lives into account demands a certain degree of foresight. In fact, foresight is needed to address most, if not all, concerns that impact society's future, such as climate change, general-purpose AI systems, soil erosion, and (engineered) pandemics. While some government agencies and departments, as well as some international organizations, are indeed tasked with the handling of a subset of these issues, many of these struggle due to their set up, scope and structure (van Aaken, 2015) to take a distinctly longer view. Taking the far future into account would require accepting and managing more imminent tradeoffs and applying foresight to all internal and external decision-making processes.

When it comes to AI, governments — by virtue of functioning on the basis of predominantly reactive policy to AI incidents and the political voting cycle — largely focus on imminent 5–10 year increments. AI incidents occurring now are

indeed urgent and should be addressed and safeguarded against. However, by extrapolation of the capability and ubiquity of these systems in past incidents, it is likely that potential future AI incidents could have an outsized negative effect on society in comparison. Indeed, the pace of development and difficulty of capturing and establishing appropriate frameworks in this space has become strikingly evident in the recent political discussions between the European Parliament, the Council of the European Union and the European Commission on the European Commission's proposal for a regulatory Act on AI (European Commission, 2021). While the initial regulatory proposal contained no reference whatsoever to general-purpose AI, various political actors have since added and amended Articles on general-purpose AI systems. This led to an animated, difficult and – at the writing of this thesis ongoing – discussion between academics,[6] civil society organizations (ALLAI, 2021) and regulators as to how these systems could and should be captured, accounting for their current (known) and eventual (unknown) capabilities.

That is precisely why it matters whether AI governance is ethically informed, future-oriented and resilient. *But how difficult could this task really be?* To put it in context, let us briefly consider another overarching and pressing area where various governance efforts have been attempted and failed throughout the past decade.

The movement for the rights of future generations[7] has, in recent years, undertaken efforts to set up governance frameworks to ensure longer-term thinking within various areas of policymaking and regulation. This includes areas that have become imminently pressing even for this generation such as climate change, where, for example, Germany's Federal Constitutional Court recently ruled that the

[6] See the discussion between Stuart Russell and Max Tegmark on the AI Act and general-purpose AI systems:
https://www.facebook.com/futureoflifeinstitute/videos/future-proofing-the-eu-ai-act-eu-parliament-hearing-highlights/1966144016926247/.
[7] See for example: https://www.appgfuturegenerations.com/.

government's climate protection mechanisms are insufficient to protect future generations.[8] This movement advocates for the development and implementation of robust governance frameworks to safeguard future generations' wellbeing, speaking on behalf of those who cannot speak for themselves and those who are already excluded by the shorter-term and narrow thinking of governments. In that sense, it is comparable to the process of establishing AI governance frameworks from scratch, accounting for known, unknown and future risks to society and individuals. Both movements aim to adjust existing governance methods or developing new ones so that an uncertain future is more likely to be net positive than not.

The movement for the rights of future generations provides an interesting case study, underscoring the difficulty of incorporating longer-term considerations into governance. It also underscores the urgency of starting this work sooner than later, and the variety of potential approaches. Many of the goals of the two areas are overlapping, such as: (i) the inclusion of both shorter- and longer-term concerns — and their respective trade-offs — in governance efforts, (ii) the inclusion and consideration of various individuals and groups that are disproportionately affected by these concerns and (iii) institution-building for novel governance measures. Moreover, the history of advocacy for future generations demonstrates the importance of thinking these concerns through *now* and, therefore, the relevance and timeliness of the research outlined in this thesis.

It may be difficult, if not impossible, to change the structures and path dependencies that governments are setting up now at a later stage. By way of comparison, there have been multiple efforts over the past decade to 'tag on' mechanisms into existing infrastructures to start taking future generations into account in governance and policy decisions. However, a lot of these efforts have been short-lived, possibly because they were poorly conceived but more plausibly because they were proposed too late, and lost to the existing institutional infrastructure,

[8] See: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2021/bvg21-031.html.

power distributions and, therefore, ingrained mechanisms. Among these, several efforts have lasted short of one election cycle. A striking example is the case of the Hungarian Commissioner for Future Generations whose independent role was discontinued after four years by the new political parties in power. Bigger efforts have faced insurmountable issues too. The Israel Commission for Future Generations, for example, was disestablished shortly after their five-year mandate, apparently due a lack of budget (Göpel & Pearce, 2013). At the time of this writing, the movement for the rights of future generations has made only two interventions whose effects have (so far) lasted. It helped establish the UK's APPG on Future Generations and the Welsh Commissioner for Future Generations. These interventions' actual impact on policymaking remains uncertain, however.

Without belaboring these attempts, it seems reasonable to assume that once a governance framework is in place, it can be difficult to amend, adjust or add governance mechanisms. Therefore, and in light of existing developments from a variety of international governments, thinking through informed, agile and future-oriented AI governance efforts matter now.

3. The European Union's lead role in the conversation about AI governance

I focus on the European Union (EU) throughout this thesis. The main reason for this is that the EU is one of the main governmental actors for which a coherent narrative from ethical considerations to government measures, such as regulation can be drawn. This is explored in Chapters II and III. The likelihood that the EU's measures will affect international decision making is explored in Chapter I. The EU has a history of successfully embedding values in governance (see e.g. the precautionary principle[9]). The Brussels effect (Bradford, 2020) demonstrates that

[9] See: https://eur-lex.europa.eu/EN/legal-content/summary/the-precautionary-principle.html.

the EU is capable of influencing external non-EU actors to adopt its laws and policies. A salient example of this is the European General Data Protection Regulation, which California emulated when it passed its own version.[10] While it is uncertain to what degree the AI Act, and, therefore, the EU will influence other, non-EU governments on the topic of AI, there are some indicators. For example, there is an increasing number of existing work streams between the US and the EU which will be influential in co-shaping a mutual vision of AI regulation and are aligned with many aspects proposed in the AI Act itself. More concretely, the Tech and Trade Council between the US and the EU will host working groups on standardization efforts towards trustworthy AI, which seems aligned with the overall ambition of the EU's AI Act. Moreover, it is noteworthy that the US NIST AI Risk Management Framework[11] strongly matches areas described in the conformity assessment of the AI Act. This combination of timeliness, potential impact and coherent strategy, aligned with the ambition of this thesis, resulted in the EU being the most promising governmental actor to focus on.

4. Overview of chapters

The chapters of this thesis have all been published as peer-reviewed papers or book chapters throughout the past 3 years of this dissertation. They are rooted in three areas: establishing the current landscape, analysis of existing approaches, and guidance for ongoing and future efforts. Given the difficulty of predicting the future capabilities of AI, and its corresponding effects, the next best thing is to develop generally robust response capabilities that can withstand some number of unexpected outcomes. This thesis explores a number of response capabilities, and, together, they form one approach to developing a framework for informed decision making on AI governance, in light of rapid technological change and uncertainty.

[10] See: https://techcrunch.com/2019/11/14/californias-new-data-privacy-law-brings-u-s-closer-to-gdpr/.
[11] See: https://www.nist.gov/itl/ai-risk-management-framework.

The stages each chapter in this thesis puts forward build on top of each other, from micro- to macro-level investigation. Each chapter put forward in this thesis functions as an individual building block that can be implemented in existing or future governance frameworks independently or in combination. Herein, I examine five approaches, perspectives and their resulting contributions towards achieving ethically informed AI governance. In the following section, I briefly summarize the core ideas, before presenting them in more depth in their dedicated chapters. The core proposals relate to each other and build on one another. First, I explore the opportunities, concerns and impact surrounding the conceptualization of 'trustworthy AI' to talk about AI ethics in governance and define the field (Chapter I). Following that, I review the larger context in which AI ethics efforts both arose within and directly informed the EU's policy-making and legal process (Chapter II). Building on Chapter I and Chapter II, I review and propose novel approaches towards ensuring that AI ethics principles, as developed across hundreds of institutions and governments in the past years (Corrêa et al., 2022; Schiff et al., 2021), can be better and more efficiently operationalized in AI governance measures (Chapter III). Next, I look at existing AI research communities operating under different AI timelines and their divergent opinions on AI governance measures, putting forward a proposal for collaboration in light of the topic's urgency (Chapter IV). Finally, all of the actions, measures and policy concerns mentioned in Chapters I through IV must be 'housed' to function in the longer run. Chapter V explores the institutional landscape within which all of this is to be implemented. It reviews existing institutions and puts forward a blueprint for ensuring adaptive, agile and future-proof institutions for ethical AI governance.

Chapter I. Artificial intelligence by any other name: a brief history of the conceptualization of 'trustworthy artificial intelligence'

One of the original, core questions in the AI governance dialogue is how to delineate and encourage only the type of AI systems that should be developed and deployed. That is, how can we implement concerns surrounding the impact of AI on fundamental rights, ethics, society, law and the environment into policy discourse. *What's in a name?*

Chapter I examines the recent conceptualization of 'trustworthy artificial intelligence' as a term increasingly ubiquitous in policy and governance dialogue. In doing so, it proposes that AI ethics and AI governance efforts are deeply intertwined. Specifically, Chapter I explores the conceptualization of the EU's approach towards achieving ethically informed AI and the impact this has had internationally. It sets the scene for normative claims and prioritizations presented in subsequent chapters.

The chapter begins by contextualizing the term, briefly investigating other, similar terminologies that have been used by governments and academics, such as 'AI for Good' or 'Beneficial AI' (Cowls et al., 2021; O'Keefe et al., 2020). Next, I explore how ethical considerations have informed the concept of 'trustworthy AI' and in doing so contributed to a unique perspective. By providing this background, the chapter outlines (i) the strategic factors that led the EU to adopt this term and (ii) the precise definition and range of concrete concepts that the term builds upon. Next, I critically examine the term. I raise several concerns about the term (such as conceptual conflations), before demonstrating the memetic appeal this initial conceptualization had when it comes to international dialogue on AI governance. In fact, it has built a first bridge between AI ethics and AI governance. I review more

than a dozen policy documents to track the term's impact.

Chapter I introduces the reader to the author's concerns about the term, its proliferation, and its potential misuse. It also primes an understanding of the broader space within which this thesis is located. It concludes with takeaways indicative of the future impact of the EU in this sphere, where I propose that it is likely, given the investigation in this chapter, that the EU will have an outsized impact on international considerations surrounding AI governance.

Most notably, the suggested impact of the EU's actions in this space has been corroborated by a new report which focuses on the EU's AI Act (Siegmann. & Anderljung, 2022) and the likelihood that it will diffuse across international governance making. Given that this regulatory framework is directly informed by 'trustworthy AI', as explained in Chapter I and further explored in Chapter II, this may contribute to a dissemination of binding and ethically informed AI governance at an international level.

Chapter II. The ghosts of AI governance past, present and future: AI governance in the European Union

The conventional wisdom is that the main players in the international policy context surrounding AI are the US and China. But given the EU's track record in the protection of individuals' rights, its current efforts towards AI regulation and the beginning of what could be a Brussels effect (Bradford, 2020) on AI regulation, this chapter poses and answers the question: *but what about the EU (Brattberg et al., 2020; Cath et al., 2018)?* In doing so, it also complements the review of the memetic impact of the EU's "trustworthy AI" concept on international governance, as presented in Chapter I.

Chapter II provides an in-depth review of the EU's approach to ethical AI governance. This serves to (i) demonstrate why the EU has been chosen as the main governmental actor within this thesis and (ii) provide a comprehensive overview of previous efforts before focusing on three specific interventions in the subsequent chapters.

The chapter undertakes a comprehensive review of the past, present and future of AI governance in the EU, weaving together policy and legal texts from a range of key actors. In doing so, it demonstrates the EU's coherent and expansive approach to ethically informed AI governance. In short, the EU ensures and encourages ethical, trustworthy and reliable technological development. It has done so in the past and there are sufficient indicators to suggest that it will continue to do so. The review covers a range of key documents and policy tools that lead to the arguably most crucial effort of the EU to date: to regulate AI with all its implications (Neuwirth, 2022; Veale & Zuiderveen Borgesius, 2021). In closing, the chapter highlights the EU's drive towards digital sovereignty through the lens of both regulation and infrastructure. It concludes by offering several preliminary considerations to achieve good and ethical AI governance in the EU. Among them are: AI megaprojects and lighthouses, AI agencies and standards. Since the writing of this chapter, all three of the identified areas have become concrete topics of discussion in academic, industry and policy discourse.

For example: a lighthouse for "safe and secure AI" has been established by a large cohort of research and academic institutes across the EU, the UK and Switzerland;[12] the establishment of a forward-looking AI Agency-model with a particular focus on general-purpose AI systems has been proposed by Members of the European Parliament Pernando Barrena Arza and Cornelia Ernst **(**"Navigator

[12] See: https://cispa.de/en/elsa.

Programme for General Purpose AI");[13] and the European Telecommunications Standards Institute appears to have recommended that standardization organizations should be tasked with the technical definition of AI, as well as the description of a categorization framework for high-risk AI, in a letter to the European Commission on the AI Act.[14]

Chapter III. Actionable principles for artificial intelligence policy: three pathways

What could be done in order to more meaningfully operationalize AI ethics principles with an eye to AI governance? In the development of governmental policy for AI, the most common sector-agnostic avenue that has been pursued is drawing up AI ethics principles. These are informed by the values, ethics and relevant lived experience of the group developing them (Adamson et al., 2019; Jobin et al., 2019). However, these AI ethics principles often fail to be implemented in governmental policy, despite a staggering 700 international policy initiatives to date[15] with over 200 international AI ethics principles.

Chapter III explores and proposes a novel framework for the development of 'Actionable Principles for AI'. My approach acknowledges the relevance of AI ethics principles and homes in on methodological elements to increase their practical implementability in policy processes. The investigation and lessons learned are important to (a) clarify how the existing work on AI ethics can be better operationalized in the increasing number of governance efforts and (b) benefit from existing expertise. Moreover, the model of an expert group, which is investigated in

[13] See: Amendment 2286 of the tabled amendments. See here: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.
[14] See: https://www.google.com/url?q=https://www.euractiv.com/section/digital/news/standardisation-body-calls-for-ai-definition-categorisation-to-be-decided-as-standards/&sa=D&source=docs&ust=1660646478553021&usg=AOvVaw11rTRTHqlwIeoiFJcYOIKs.
[15] See: https://oecd.ai/en/dashboards.

this chapter, provides insights into governance mechanisms that arose after the publication of this chapter as a stand alone paper in AI and Ethics in 2021. For example, it can serve to provide relevant considerations for (a) transatlantic expert groups — which will involve technical, governance and ethical talent — such as in the US-EU Tech and Trade Council and (b) European working groups with a similar mix of key expertises as envisioned for the working groups supporting the upcoming European AI Board.[16]

The chapter makes use of the most impactful expert group for AI ethics and policy in the EU to date: the European Commission's High Level Expert Group on AI (AI HLEG). It evaluates their working processes for developing the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) to extract relevant procedural lessons. Subsequently, the working processes are evaluated in light of their ability to contribute to a prototype framework for the development of 'Actionable Principles for AI'. The chapter also reviews several shortcomings of the work of the AI HLEG. As a result of this chapter's investigation, I propose the following three components of forming such a prototype framework: (1) preliminary landscape assessments; (2) multi-stakeholder participation and cross-sectoral feedback; and, (3) mechanisms to support implementation and operationalizability. In doing so, Chapter III also complements investigations undertaken and conclusions drawn in Chapters I and II by virtue of proposing supplementary processes to ensure that ethical approaches to AI governance such as those championed by 'trustworthy AI' become more implementable and concrete.

Since the publication of this chapter, academic research has increasingly focused on specific ethical concerns outlined in AI ethics principles and how these can be technically implemented, with a promising rise in discussions of 'audits' and their implications (Costanza-Chock et al., 2022; Raji et al., 2022). Similarly, policy

[16] The European AI Board will be supporting the development and implementation of the AI Act and will be advised by a range of expert working groups as per Art. 57 and 58 in COM/2021/206 final.

discourse has heavily drawn on AI ethics principles, especially in the EU. As projected in this chapter, the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) have directly inspired law-making. Not only do several ethical requirements closely match the legal obligations outlined in the AI Act (European Commission 2020), the second compromise text proposed by the Czech presidency of the Council of the European Union to the AI Act[17] directly references the influence of the content of the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) and the AI HLEG on new legal provisions, which will ultimately directly impact AI development and deployment in the EU:

"More specifically, Article 6(3) has been extended and it now contains new provisions inspired by ideas from the AI HLEG and from the OECD classification framework of AI systems, according to which the significance of the output of the AI system in relation to the decision or action taken by a human, as well as the immediacy of the effect should also be taken into account when classifying AI systems as high risk."[18]

A 'general-approach' has been reached in the Council of the European Union since, and this paragraph has been implemented and removed throughout several iterations.[19]

Chapter IV. The case for an 'incompletely theorized agreement' on AI governance

Given the goal of ensuring that AI doesn't negatively affect society, now and in the

[17] See the Czech document dated 15/07/2022 here: https://www.kaizenner.eu/post/aiact-part3.
[18] An exploration of the impact of the AI HLEG's work on the OECD's work (as referred to here) can be found in Chapter I. I argue that there is significant reason to believe that the OECD's work was shaped by the AI HLEG.
[19] See the final 'general-approach' here:
https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

distant future, *how can advocates of the near-term and longer-term future work together to achieve the overarching goal of developing 'trustworthy AI'* ?

It is clear that recent progress in AI raises a wide array of ethical and societal concerns. Accordingly, a range of appropriate policy interventions are needed and, in some cases, developed by governments, as explored in Chapters II and III. While there has been a wave of scholarship in this field, Chapter IV proposes that the research community at times appears divided between those who emphasize 'near-term' concerns and those focusing on 'long-term' concerns and corresponding policy measures. This increasing tension — and the associated policy, financial and research impacts — merits a deeper exploration. Ideally, those concerned with the near-term and far-term could collaborate meaningfully on their shared higher-level goal.

In Chapter IV, I work with a co-author to map and critically examine this apparent 'gulf', with a view to understanding the practical space for inter-community collaboration on AI policy. There is an increasing sense that premature fragmentation of this relatively nascent field can become a detriment to achieving a comprehensive, coherent and appropriate approach to governing AI. In particular, given the stakes, lessening fragmentation where it isn't vital appears to be an important workstream to achieve an overarching framework to safely develop and deploy AI for the benefit of all. This chapter culminates in a proposal to make use of the legal notion of an 'incompletely theorized agreement'. In using this notion, the chapter suggests that on certain issue areas, scholars working with near-term and long-term perspectives can converge and cooperate on selected, mutually beneficial AI policy projects, all the while maintaining divergent perspectives.

This chapter was inspired by the underlying worry that the future of AI governance, as a field, relies on a variety of scholars (e.g. technical, ethical and legal experts) and policy actors (e.g. regulators) who may have strongly divergent opinions as to

what should be done. Unfortunately, this concern has been substantiated in recent times. A wide range of very public disagreements that appear to be unsettled, with parties creating increasingly entrenched positions of opposition regarding research, funding and direction of the field have surfaced in classic media,[20] blogs,[21] and on social media.[22]

Chapter V. Foundations for the future: institution building for the purpose of artificial intelligence governance

As governance efforts for AI are becoming increasingly concrete, it is becoming increasingly crucial to draw on a variety of approaches and instruments. These include hard regulation; standardization efforts and mitigating challenges from risky AI systems, their practical coming into force and monitoring. To implement these and other efforts, new institutions will need to be established on a national and international level.

Chapter V draws on the lessons learned throughout the previous chapters and builds towards a bigger picture. Concretely, it examines how new institutional frameworks can be established so as to ensure that AI governance mechanisms (such as those explored and proposed in Chapters I through IV) are possible and will function well. The chapter proposes blueprints for various institution building scenarios and investigates three key components of any future AI governance institutions, exploring the benefits and drawbacks of each. In particular, it examines: (1) "purpose," relating to the institution's overall goals and scope of work or mandate); (2) "geography," including questions of participation and the reach of

[20] See: https://www.vox.com/future-perfect/2022/8/10/23298108/ai-dangers-ethics-alignment-present-future-risk.
[21] See: https://spectrum.ieee.org/timnit-gebru-dair-ai-ethics; and discussions here: https://www.lesswrong.com/posts/R3tXGhSCgYbp3kXm2/jack-clark-s-spicy-ai-policy-takes.
[22] See: https://twitter.com/jackclarkSF/status/1555980661499908096; and here: https://twitter.com/timnitGebru/status/1486093692741980160.

jurisdiction and (3) "capacity," which depends on each institution's infrastructure and staff. Subsequently, it highlights noteworthy aspects of various institutional roles, with a focus on "purpose." In order to explore what these proposals could look like in practice, the chapter concludes by placing these debates in a European context and proposing different iterations of a potential European AI Agency. In doing so, it also builds on several predictions Chapter II made in its conclusion.

Novel efforts to achieve bilateral cooperation have been launched. For example, there is the Tech and Trade Council between the US and the EU, the nascent international cooperation on AI through the OECD's AI observatory and international fora, such as the new working groups via the Global Partnership on AI. These networks have shaped how we envision new institutional mechanisms that encourage safe and beneficial cross-border AI development and deployment, as well as our belief as to the efficiency with which we expect them to function and implement concrete actions.

The real and felt importance of suitable institution building can be equally found in recent academic publications, such as in the Stanford Human-Centered AI Institute's white paper (Zhang et. al, 2022) on building a Multilateral AI Research Institute (MAIRI) in the US, to boost the US's leadership for AI research and AI governance models. A similar model to what a European AI Agency could achieve, as investigated in Chapter V, has been recently proposed under the concept of an 'AI Control Council', for the US context (Korinek, 2021).

References

Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems. *Proceedings of the IEEE*, *107*(3), 518–525. https://doi.org/10.1109/JPROC.2018.2884923.

AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. European Commission, High Level Expert Group on AI.https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

ALLAI. (2021). Artificial Intelligence Act: Analysis and Recommendation. https://allai.nl/wp-content/uploads/2021/08/EU-Proposal-for-Artificial-Intelligence-Act-Analysis-and-Recommendations.pdf.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1606.06565.

Anderson, J. M., Kalra, N., Stanley, K., Sorensen, P., Samaras, C., & Oluwatola, T. A. (2016). *Autonomous Vehicle Technology: a Guide for Policymakers*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR443-2.html.

Askell, A., Brundage, M., & Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1907.04534.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 004912411878253. https://doi.org/10.1177/0049124118782533.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-60648-4.

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. https://play.google.com/store/books/details?id=mZXHDwAAQBAJ.

Brattberg, E., Rugova, V., & Csernatoni, R. (2020). *Europe and AI: Leading, lagging behind, or carving its own way?* (Vol. 9). Carnegie Endowment for International Peace Washington. https://carnegieendowment.org/files/BrattbergCsernatoniRugova_-_Europe_AI.pdf.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1802.07228.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2004.07213.

Bryson., J. (2018). No one should trust artificial intelligence. *Our World: Brought to you by United Nations University*.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 8*(1), 77–91. PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html.

Buyers, J. C. (2018). *Artificial intelligence: the practical legal issues*. Law Brief Publishing.

Calo, R. (2010). Peeping HALs: Making Sense of Artificial Intelligence and Privacy. *EJLS - European Journal of Legal Studies*, *2*(3), 168–192. http://www.ejls.eu/6/83UK.htm.

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, *51*, 399. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/davlr51&section=18

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0080.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the "good society": The US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7.

Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(1753). https://papers.ssrn.com/abstract=3213954.

Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books. https://play.google.com/store/books/details?id=TdL2DwAAQBAJ.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of Humanity Institute. University of Oxford*. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Cihon, P., Maas, M. M., & Kemp, L. (2020). Should Artificial Intelligence Governance be Centralised? Design Lessons from History. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2001.03573.

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information. An International Interdisciplinary Journal*, *12*(7), 275. https://doi.org/10.3390/info12070275.

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press. https://play.google.com/store/books/details?id=Gs_XDwAAQBAJ.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., & Terem, E. (2022). Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2206.11922.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213.

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, *3*(2), 111–115. https://doi.org/10.1038/s42256-021-00296-0.

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://play.google.com/store/books/details?id=KfodEAAAQBAJ.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., Raji, D., Rankin, J. L., Richardson, R., Schultz, J., West, S. M., & Whittaker, M. (2019). *AI Now 2019 Report*, AI Now Institute, 100. https://ainowinstitute.org/AI_Now_2019_Report.pdf.

Cremer, C. Z., & Whittlestone, J. (2021). Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI. In *International Journal of*

Interactive Multimedia and Artificial Intelligence, 6(5), 100. https://doi.org/10.9781/ijimai.2021.02.011.

Danaher, J. (2019). *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press. https://play.google.com/store/books/details?id=NX6mDwAAQBAJ.

Danzig, R. (2017). An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order. *Lawfare Research Paper Series*, *5*(1). https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf.

Deeks, A. (2020). High-Tech International Law. *The George Washington Law Review*, *88*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531976&download=yes.

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *arXiv:1710.00794 [cs]*. http://arxiv.org/abs/1710.00794.

Ess, C. (2006). Ethical pluralism and global information ethics. *Ethics and Information Technology*, *8*(4), 215–226. https://doi.org/10.1007/s10676-006-9113-3.

European Commission. (2020). *Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act)* (COM/2020/767 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767.

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., Potgieter, P., Vu, K., Whalley, J., & Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy*, *44*(6), 101988. https://doi.org/10.1016/j.telpol.2020.101988.

Feldstein, S. (2019). The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression. *Journal of Democracy*, *30*(1), 40–52. https://doi.org/10.1353/jod.2019.0003.

Fischer, S.C., Leung, J., Anderljung, M., O'keefe, C., Torges, S., Khan, S. M., Garfinkel, B., Dafoe, A., Brundage, M., Rey Ding, J., Flynn, C., Maas, M., Matheny, J., Daniel, M., Lintz, A., Muehlhauser, L., Page, M., Shevlane, T., Toner, H., … Gar, B. (2021). *AI policy levers: A review of the U.S. government's tools to shape AI research, development, and deployment*. Retrieved June 1, 2022, from https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deploym

ent-%E2%80%93-Fischer-et-al.pdf.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. https://doi.org/10.2139/ssrn.3518482.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., & Others. (2018). AI4People White Paper: Twenty Recommendations for an Ethical Framework for a Good AI Society. *Forthcoming in Minds and Machines, December*.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Garcia, D. (2018). Lethal Artificial Intelligence and Change: The Future of International Peace and Security. *International Studies Review*, *20*(2), 334–341. https://doi.org/10.1093/isr/viy029.

Gasser, U. (2016). *Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy*. Harvard Law Review Forum, Law, Privacy & Technology Commentary Series, 130(December), 10. https://harvardlawreview.org/2016/12/recoding-privacy-law-reflections-on-the-future-relationship-among-law-technology-and-privacy/.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *The Journal of Artificial Intelligence Research*, *62*, 729–754. http://www.jair.org/index.php/jair/article/view/11222.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://scholarspace.manoa.hawaii.edu/handle/10125/59651.

Greene, K. G. (2022). AI Governance Multi-stakeholder Convening, in Bullock, J.B. et al. (eds), *The Oxford Handbook of AI Governance*. https://doi.org/10.1093/oxfordhb/9780197579329.013.6.

Gruetzemacher, R., & Whittlestone, J. (2019). Defining and Unpacking Transformative AI. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1912.00747.

Gutierrez, C. I., & Marchant, G. E. (2021). *A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence*. https://doi.org/10.2139/ssrn.3855171.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American*. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2019). Will Democracy Survive Big Data and Artificial Intelligence? In Helbing, D. (ed.), *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*, Springer International Publishing, 73–98. https://doi.org/10.1007/978-3-319-90869-4_7.

Horowitz, M. C. (2018, May 15). *Artificial Intelligence, International Competition, and the Balance of Power*. Texas National Security Review. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174.

Horowitz, M. C., Scharre, P., & Velez-Green, A. (2019). A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence. *arXiv:1912.05291 [cs]*. http://arxiv.org/abs/1912.05291.

Jelinek, T., Wallach, W., & Kerimi, D. (2020). Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-020-00019-y.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., & Choi, Y. (2021). Delphi: Towards Machine Ethics and Norms. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2110.07574.

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1906.11668.

King, T., Aggarwal, N., Taddeo, M., & Floridi, L. (2018). *Artificial Intelligence*

Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. https://doi.org/10.2139/ssrn.3183238.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, *26*(1), 89–120. https://doi.org/10.1007/s11948-018-00081-0.

Korinek, A. (2021). *Why we need a new agency to regulate advanced artificial intelligence: Lessons on AI control from the Facebook Files*. The Brookings Institution. https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/.

Lewis, D. A., Blum, G., & Modirzadeh, N. K. (2016). *War-Algorithm Accountability* (ID 2832734). Social Science Research Network. http://papers.ssrn.com/abstract=2832734.

Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press. https://books.google.nl/books?id=oBb-lt3l4oYC.

Livingston, S., & Risse, M. (2019). The Future Impact of Artificial Intelligence on Humans and Human Rights. *Ethics & International Affairs*, *33*(2), 141–158. https://doi.org/10.1017/S089267941900011X.

Maas, M. M. (2019). International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order. *Melbourne Journal of International Law*, *20*(1), 29–56. https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3144308/Maas.pdf.

Maas, M. M. (2021). *Aligning AI Regulation to Sociotechnical Change*. https://doi.org/10.2139/ssrn.3871635.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*.

https://doi.org/10.1007/s11948-019-00165-5.

Müller, V. C. (2020a). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Müller, V. C. (2020b). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. CSLI, Stanford University. https://plato.stanford.edu/entries/ethics-ai/.

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In Müller, V.C. (Ed.), *Fundamental Issues of Artificial Intelligence*. Springer International Publishing, 555–572. https://doi.org/10.1007/978-3-319-26485-1_33.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0089.

Neuwirth, R. J. (2022). *The EU Artificial Intelligence Act: Regulating Subliminal AI Systems*. Taylor & Francis. https://play.google.com/store/books/details?id=zBN1EAAAQBAJ.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://play.google.com/store/books/details?id=-ThDDwAAQBAJ.

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice: An International Forum*, *19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2.

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020). The Windfall Clause: Distributing the Benefits of AI for the Common Good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*). Association for Computing Machinery, 327–331. https://doi.org/10.1145/3375627.3375842.

Pennachin, C., & Goertzel, B. (2007). Contemporary Approaches to Artificial General Intelligence. In *Artificial General Intelligence*,1–30. https://doi.org/10.1007/978-3-540-68677-4_1.

Prunkl, C. E. A., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, *3*(2), 104–110.

https://doi.org/10.1038/s42256-021-00298-y.

Prunkl, C., & Whittlestone, J. (2020). Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 138–143. https://doi.org/10.1145/3375627.3375803.

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. https://doi.org/10.1145/3514094.3534181.

Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks*. https://doi.org/10.2139/ssrn.3259344.

Roberts, H., Cowls, J., Hine, E., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). *Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes*. https://papers.ssrn.com/abstract=3811034.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2019). Tackling Climate Change with Machine Learning. *arXiv:1906. 05433 [cs, Stat]*. http://arxiv.org/abs/1906.05433.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin. http://aima.cs.berkeley.edu/~russell/papers/mi19book-hcai.pdf.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61–86. https://doi.org/10.1108/JICES-12-2019-0138.

Samuel, A. L. (1960). Some Moral and Technical Consequences of Automation—A Refutation. *Science*, *132*(3429), 741–742. https://doi.org/10.1126/science.132.3429.741.

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. https://doi.org/10.1145/3375627.3375804.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI Ethics in the Public,

Private, and NGO Sectors: A Review of a Global Document Collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42. https://doi.org/10.1109/TTS.2021.3052127.

Siegmann, C., Anderljung, M. (2022). The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market. https://uploads-ssl.webflow.com/614b70a71b9f71c9c240c7a7/62fbe1c37eff7d304f0803ac_Brussels_Effect_GovAI.pdf.

Stanley, J. (2019). *The Dawn of robot surveillance: AI, Video analytics, and privacy*. American Civil Liberties Union.

Tasioulas, J. (2019). First Steps Towards an Ethics of Robots and Artificial Intelligence. *Journal of Practical Ethics*, *7*(1), 61–95. http://www.jpe.ox.ac.uk/papers/first-steps-towards-an-ethics-of-robots-and-artificial-intelligence/.

Turner, J. (2018). *Robot Rules: Regulating Artificial Intelligence*. Springer. https://play.google.com/store/books/details?id=moB1DwAAQBAJ.

Zhang, D., Larence, C., Sellitto, M., Wald, R., Schaake, M., Ho, D., Altman, R., Grotto, A.(2022). Enhancing International Cooperation in AI Research: The Case for a Multilateral AI Research Institute. Stanford Human Centered Artificial Intelligence.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9.

van Aaken, A. (2015). *Is International Law Conducive to Prevent Looming Disasters?* University of St. Gallen Law School. https://play.google.com/store/books/details?id=qvX-vQEACAAJ.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act. In *SocArXiv*. arXiv. https://doi.org/10.31235/osf.io/38p5f.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M., & Nerini, F. F. (2019). The role of artificial intelligence in achieving the Sustainable Development Goals. *arXiv:1905.00501 [cs]*. http://arxiv.org/abs/1905.00501.

Walz, & Firth-Butterfield. (2019). Implementing ethics into artificial intelligence: A contribution, from a legal perspective, to the development of an AI governance regime. *Duke Law and Technology Review, 18*, 176–231. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/dltr17&section=10.

Whittlestone, J., & Ovadya, A. (2019). The tension between openness and prudence in AI research. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1910.01170.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*(3410), 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Chapter I

Artificial intelligence by any other name: A brief history of the conceptualization of "Trustworthy Artificial Intelligence"

……..

Forthcoming: Collection on National AI Strategies, *Discover Artificial Intelligence*, Springer.

……..

Introduction

Recent years have seen an increase in artificial intelligence (AI) capabilities and incidents. Correspondingly, there has been an influx of government strategies, panels, dialogues and policy papers, including efforts to regulate and standardize AI systems (Fischer et al., 2021; Dafoe, 2020; Brundage et al., 2020; Maas, 2021). A first step in most of these efforts is to delineate the scope of the resulting document, typically by either outlining a range of standard technical definitions of AI (Wang, 2019; Samoili et al., 2020) or referencing existing scholarly work (Russell & Norvig, 1995). After defining their scope, many policy documents published by governments delve deeper into the 'type' of AI they wish to solicit from industry players and deploy nationally or globally. This largely serves to ensure that the strategies, policy discussions and AI-related milestones sketched within these documents are guided by a 'north star', or overarching goal. The north star should be comprehensible to all who read and implement the document. Describing the north star allows a

non-technical audience to follow and partake in the relevant policy discussions, though it does not replace technical definitions. Although more could be said as to *why* this is being done and whether it is sensible, such discussion is outside the scope of this paper. Instead, I focus on and contextualize some of these 'north star' definitions themselves. In particular, I explore one of the most prominent recent descriptions: the EU's concept of "trustworthy AI." I explain its background, its international effects and its drawbacks in more depth. *What is in a name?* What is in *"trustworthy AI*?"

1. Other terms

To provide proper context, this section describes a number of terminologies that political decision makers have used to reference the type of AI they desire to encourage. This serves two purposes. First, it primes the subsequent investigation of the term "trustworthy AI." Second, it highlights the difficulty of choosing and solidifying an appropriate term for use in governmental contexts. Below is a non-comprehensive selection of some of the most prominently used terms within the past couple of years. It should be noted that different groups are responsible for coining and/or advocating for each term. This is to say that these terms have not necessarily been originated by governments, though they have been picked up by governmental discourse. Moreover, the terms refer to varying objectives and measures as to how to achieve those objectives. A commonality across all these terms is that they look to describe and capture AI systems that will bring benefits to society — those that presumably will make the world a better place in the near future and for future generations.

1.1. Ethical AI

Following a wave of AI ethics documents, charters and public AI ethics discussions (Schiff et al., 2020; Ryan & Stahl, 2020; Hagerty & Rubinov, 2019), one increasingly popular term is "**ethical AI**."

The term "ethical AI" has been widely used to refer to AI systems that are in line with our moral values (Christian, 2021; Coeckelbergh, 2020; Müller, 2021). The term has had one of the earliest, strongest and most continuous influences on public and governmental discourse. The nascent field of AI ethics is increasingly crucial as we tackle the manifold potential harms society has suffered by AI systems — and as we work to preempt potential harms. For example, the following concerns have deeply impacted various groups in recent times: algorithmic bias (Barocas & Selbst, 2016; Crawford et al., 2019; Berk et al., 2018); transparency and explainability (Doran et al., 2017; Gebru et al., 2021); the safety of autonomous vehicles (Anderson et al., 2016; Nyholm & Smids, 2016); privacy concerns in the face of widespread surveillance (Calo, 2010; Gasser, 2016), and the economic and political effects of technological unemployment (Frey & Osborne, 2017; Danaher, 2019).

It should be noted that, while most interpret "ethical AI" to refer to AI that is in line with at least a subset of common ethical considerations,[23] some may (mis)interpret it as AI that exhibits 'ethical' behavior and, by extension, is a moral agent.[24] Having said that, it is a challenge to explore "ethical AI" as a concept per se. The term's ambiguity simultaneously minimizes the true nature of the field and maximizes the appeal of the term, without creating responsibility within the user to clearly define it. Accordingly, the term has indeed been co-opted, especially in media and speaker circuit discourse. The large and shifting scope of the term may explain why other, equally popular terms have arisen. Presumably, there has been a need for terms that more clearly delineate suggested concepts and goals.

[23] Discussions around finding agreement on which issues are the most important and universally agreed upon, as well as shortcomings of and complexities in that process, are outside the scope of this paper.
[24] This interpretation is explicitly not explored or understood to be referenced throughout this paper.

1.2. AI for good

For example, another commonly used term, particularly from the earlier days of AI policy discourse, is "**AI for good**." This term has been particularly appealing to discourse within industry, though it has been equally popular with governmental actors. Whether the term is positive or negative in terms of strategic messaging and impact is outside the scope of this paper.

"AI for good" has become a hallmark for the United Nations International Telecommunications Unit (UN ITU) in particular. The UN has built a digital platform around the term, which is designed to encourage discussions and projects aimed at finding practical solutions for the UN's Sustainable Development Goals (SDGs) through AI (Vinuesa et al., 2020). The term "AI for good" then, in this context, refers to AI systems that help solve previously identified, complex global issues for society, thus benefiting humankind (SDG, 2015). Those working to develop AI for good measure their success by their AI systems' ability to help society reach certain goals.

1.3. Beneficial AI

Another popular term, initially promoted more by the research community than by governments, is "**Beneficial AI**."

"Beneficial AI" was spearheaded through the Future of Life Institute's Asilomar Conference on Beneficial AI. This conference also yielded one of the very first sets of AI Principles, the Asilomar AI Principles[25] which were signed by 5720 people, including 1797 AI and robotics researchers. Given the principles described in the document, the term initially appears to have had a close connection to technical AI

[25] See: https://futureoflife.org/2017/08/11/ai-principles.

research. In particular, it referred to topics that fall under the research field of AI safety (Amodei et al., 2016; Leike et al., 2017; Christiano et al., 2017.), as well as topics related to the long-term future.[26] Outside of this research space, the term has been co-opted to broadly reference AI systems that *benefit* society and the environment, while avoiding definable and undefinable harm.

1.4. Responsible AI

A fourth commonly used term is "**responsible AI**." While the previously mentioned terms refer mainly to AI systems, "responsible AI" more often refers to the actions, and actors, involved in developing and deploying those systems.

"Responsible AI" seems a more sophisticated term, perhaps due to the general connotations of "*responsibility."* It has been used to refer to many of the mechanisms or methods by which it could feasibly be achieved, such as *responsible design and development* for AI (Barredo Arrieta et al., 2020; Dignum, 2019). It seems this term most often refers to processes that result in technical achievement and meet certain standards of responsibility. Some actors, particularly industry actors, likely feel that this term is more precise than terms that invoke ethics. It may be a preferable term from a communications perspective as it, similar to "trustworthiness," references a particular kind of behavior we regard as good when it is displayed by individuals or organizations.

To a degree, all of the aforementioned terms are open to interpretation, which may be welcome (or even intended) by some actors using these terms (cf., for example, discussions of ethics washing and ethics shopping (Morley, Kinsey, et al., 2021; Morley, Elhalal, et al., 2021).

But what about "trustworthy AI?" The following section summarizes this term's

[26] See: https://www.bbvaopenmind.com/en/articles/provably-beneficial-artificial-intelligence/.

history and the scope of its original definition, demonstrating that, in principle, it is very clearly defined. Subsequently, Section 3 critically examines some downsides of the term and places its impact on the international AI governance debate in context.

2. "Trustworthy AI:" the origin story

Following the publication of its AI Strategy (European Commission, 2018a), and the aim outlined therein to put forward an ethical and regulatory framework for AI, the European Commission established an independent expert group to fulfill part of this commitment. The expert group, the High-Level Expert Group on AI (henceforth: AI HLEG), was tasked with a number of projects. Most notably and relevant to this paper, their primary task under their initial mandate was to develop ethics guidelines for AI.

The AI HLEG, composed of 52 subject experts from various sectors and fields of expertise, began a comprehensive and iterative process to establish what ultimately became a building block for AI governance in the EU. Although a majority of the work was conducted internally, the AI HLEG shared their progress in meetings open to institutional observers. They solicited feedback on their first draft of the ethics guidelines half a year into the process via the AI Alliance (Stix 2021a, Stix 2021b), a platform through which the public and institutional actors were able to interact with the AI HLEG. In that first draft (AI HLEG, 2018), the conceptualization proposed by the AI HLEG was that AI should: (1) "respect fundamental rights, applicable regulation and core principles and values, ensuring an 'ethical purpose'" and (2) "be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm." The framework outlined to achieve trustworthy AI was built around (i) ethical purpose

based on values and principles as enshrined in human rights law and other relevant charters, (ii) realization of trustworthy AI through technical and non-technical methods and (iii) an assessment list with use cases for developers, deployers and users operationalizing trustworthy AI.

Following the implementation of public feedback, the AI HLEG presented their final Ethics Guidelines for Trustworthy AI in April 2019 (AI HLEG, 2019).

The final Ethics Guidelines for Trustworthy AI proposed, for the first time, a complete conceptual understanding and agreement as to what type of AI should be encouraged within the EU. While the document is strongly anchored in EU values and fundamental rights, as enshrined in the Charter of Fundamental Rights of the European Union,[27] the core concept of trustworthy AI is novel.

Trustworthy AI, in its final form, is defined as being composed of three parts. In order for an AI system to count as "trustworthy," (1) it must be lawful; that is, adhering to all legal obligations which are binding and required at that time, (2) it should be ethical; that is, adhering to and fulfilling all ethical key requirements that have been put forward in the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019), and (3) it should be robust, both from a technical and a social perspective. The last means that it should be robust in functionality, accurate, reliable, resilient to attack and other cybersecurity and security considerations. Equally, it should be robust within society and the environment; it should support beneficial societal processes and encourage cohesion and a well-functioning society. This corresponds to pillar 2 in the draft Guidelines (AI HLEG, 2018).

Given the depth and scope of these three pillars, it is clear that the conceptualization has to do a lot of heavy lifting. After all, the second pillar alone — the ethical component — is itself composed of seven key requirements that were

[27] See: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

spelled out by the AI HLEG in the very same document. In order for an AI system to adhere to the second pillar alone, it must meet standards in the following areas: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-Discrimination and Fairness; Societal and Environmental Well-Being; and Accountability. These key requirements, as they are referred to in the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019), were themselves distilled from four core principles the AI HLEG agreed upon in correspondence with feedback from the AI Alliance and other actors: Respect for Human Autonomy, Prevention of Harm, Fairness and Explicability.

The three pillars take a lifecycle approach, requiring monitoring and adherence from the research stage to deployment and long after, even as new standards and regulations are developed. They also impose a range of active prescriptive requirements on a range of different actors, such as technical researchers, governments and users.

This paper's aim is not to evaluate to what degree these pillars are sufficient or, indeed, the best ones to use as reference points (see further work on this by e.g. Mökander et al., 2022; Salo-Pöntinen & Saariluoma, 2022). Rather, it looks to demonstrate that there has been a proliferation of terms similar to "trustworthy AI" in government dialogue, and that the EU was the first (and so far, only) governmental actor that has delineated its preferred term in terms of clear, identifiable, and verifiable requirements.[28] Looking ahead, the conceptual clarity and coherence of "trustworthy AI" within the work of the EU, as well as its wide adoption, may provide some insight as to the EU's influence over international AI policy making.

[28] It should be mentioned that, outside of governmental dialogue, various AI-relevant research communities have developed multiple frameworks and conceptualizations. Indeed, as the AI HLEG itself was composed of various experts from diverging research fields, the conceptualization of trustworthy AI was certainly inspired by, and building on, those research communities' work.

3. "Trust in AI," "trusted AI" or "trustworthy AI:" Promises, perils and problems

While the term "trustworthy AI" now refers to a clear enumeration of values, rights and technical specifications, the word "trustworthy" may itself be reason for some concern. "Trustworthy" is, arguably, extremely open to interpretation and, as such, likely to cause unwitting confusion, or to be misinterpreted or misused, intentionally or not. Unfortunately, this ambiguity could undermine the clear definition (and the crisp intention and requirements) it is supposed to provide, as investigated in 3.1.a. The risk of misinterpretation or misuse is even higher when "trustworthy AI" is reused, amended and adopted in international policy contexts *without the original relation to the three pillars and their corresponding requirements*, or with only partial reference to them. The wide adoption of the term runs the risk of diluting its meaning beyond recognition and backfiring on serious policy efforts, by way of losing both the original intent and the core content the terminology is meant to encourage after all.

3.1. Come hell or waters high: Is this *trustworthy* AI?

This subsection highlights some initial concerns around the term "trustworthy AI." It proposes two overarching concerns. First, various meanings of the term, depending on actor and context, can easily get conflated, effectively creating different meanings and understandings which can mislead. Second, the term could easily be 'washed out' of its original meaning through repetition,[29] whether or not this is the underlying intent from a strategic and self-interested perspective of an

[29] Akin to the 'telephone' word repetition game in which the original sentence gets lost after sharing it too often along a chain of individuals.

agent. In short, one concern is with the word "trustworthy" itself, another with the repeated usage of the word and its associated definition. This creates the backdrop for the subsequent review in Section 3.2., which examines the dissemination of the term and how it has seemingly been re-used in various political contexts, as well as the results of that usage.

3.1.a. Conflation

The term "trustworthy AI" conflates at least five meanings:

- trust *in* the proper functioning and safety of the technology;
- the technology *being* worthy of the trust of the humans making use of it or encountering it otherwise;
- humans making use of it or encountering it *seeing* the technology as trustworthy;
- humans making use of it or encountering it *experiencing* the technology as trustworthy;
- the technology that is *worthy of trust* to all.

W*hat does "trustworthy" mean in different contexts and to different actors*? It is questionable whether any technology can even be *trustworthy*, given that this is a concept predominantly applied to human-human interactions (Bryson, 2018). More specifically, the term is typically applied to interactions in which we cannot be certain of the intentions of another person, but we assume they are innocuous based on the person's past actions and other social signals (Lockey et al., 2021). When we step onto a plane or into a car, we do not necessarily describe these technologies as "trustworthy." Instead, we rely on the fact that they have been sufficiently tested and passed all necessary thresholds to be safe for us to use (Winter et al., 2021; ÓhÉigeartaigh et al., 2020). If anything, we *trust the humans* that have been involved in ensuring these technologies are safe (Brundage et al., 2020). By

extension, we trust the institutions, organizations and processes they have set up, are involved in and are accountable to.

Calling an AI system "trustworthy" seemingly personifies the technology. Just as we may have to resort to trusting a person when we cannot know their intentions, trusting an AI system seems to imply that it has inaccessible intentions that we have no control over, and that we are fine with that status quo. This implicitly assigns human-like properties to the AI and weighs them more heavily than is desirable. Moreover, calling an AI system "trustworthy" downgrades the expectations we might rightly have about being able to access a sufficient understanding of the full complexities, mechanisms and safety implications of the AI systems we encounter, by virtue of 'assumed trust' through the AI system's trustworthiness.

Moreover, those using the term "trustworthy AI" risk conflating the meanings of *trusted* and *trustworthy*. An individual such as the Tinder Swindler (*Tinder Swindler* [documentary], Netflix 2022) appears to have been (unfoundedly) trusted but he was not actually trustworthy. As humans trusting other humans we mostly deal with incomplete sets of information about the individual. We have to infer through past actions, behaviors, social networks and adjacent signals whether engaging with an individual is safe. While the concept of trustworthy AI is intuitively appealing, "trustworthy" does not suffice in a space where we need to have a reliably and sufficiently complete set of information — either directly or validated through experts — to ensure that indeed, the technology we engage with is sufficiently safe and desirable. In particular, there is a nagging worry that the term appeals to an intrinsically human concept and strongly overemphasizes the degree to which we should *trust* an AI system without expert supervision and access to empirical facts.

Even for some experts of the AI HLEG, the range of interpretations of "trustworthy" and its relation to various actors in the ecosystem remain manifold and in need of clarification. ALLAI, an organization founded by three members of the AI HLEG, recently provided feedback on the European Commission's proposal for a horizontal regulation for AI stating that (ALLAI, 2021): "First of all, in the current wording the scoring should be aimed at evaluation or classification of trustworthiness of people. While we like the term trustworthiness for obvious reasons, in this context it is vague. What is considered the trustworthiness of a person?" (p. 11).[30] Indeed, what *is* the trustworthiness of a person? If we can't measure the trustworthiness of a person, why would we use this term to evaluate a technology, particularly when we seek to establish an evaluative framework that has clear, contextualized and verifiable metrics? The overall goal should be to encourage AI systems that function as expected, designed and deployed for, and are legal and robust. The goal is not to encourage systems that are trustworthy in the intuitive, colloquial sense. This would also match the actual work and suggestions of the AI HLEG.

Another tempting (mis)interpretation may be that an AI system is in a way responsible for its 'trustworthiness'. It should be noted that this is somewhat speculative and not reflective of any ongoing discussion in policy discourse. The speculative risk here, especially when it comes to using this term within a lay population, is that "trustworthy AI" subtly but distinctly deflects from *who* and *what* we want to trust. The terminology is capable of being interpreted, in a very subtle way, such that the responsibility to be trustworthy falls onto the AI system itself. As previously stated, presumably, what we do want to trust is that all the technical aspects of an AI system — everything that contributes to its functioning — have been adequately tested and developed. We want to trust that these technical aspects fulfill and pass all requirements necessary, that eventual misuses have been tested for and prevented to the best of the current state of the art of the

[30] The use of "trustworthiness" also appears in the proposal for a horizontal regulation for AI under 5 (1) c.

technical research available. We want to trust that the AI system has been deployed in a manner that is in line with fundamental rights, the law and ethics. Presumably *who* we actually want to trust are the researchers, the individuals involved in the entire lifecycle of the AI system, those deploying it on the market and those testing the systems for benchmarks, certifications and standards. We want to ensure that they have good intentions and that their work is accessible for third parties to verify the accuracy and intent of it. And the way to do this is to verify our trust by outsourcing it to existing methods, be that audits, employment contracts, existing regulation or otherwise.

3.1.b. Rinse and repeat: 'Washing out' any distinct meaning

Finally, there is a danger that the seductive graspability and familiarity of the term "trustworthy" could be weaponized by some actors to obfuscate the development process of their AI systems. While the EU has been incredibly clear as to what obligations an AI system's life cycle needs to fulfill for the system itself to be trustworthy, the meaning of the term is becoming more vague as it becomes increasingly popularized. Therefore, it can fall prey to being "green washed" (Ramus & Montiel, 2005) or "ethics washed" (Morley, Elhalal, et al., 2021; Bietti, 2021). The term may be misused to reassure consumers about an AI system that may not actually adhere to any of the expected guardrails or checks. There is no law surrounding what can and cannot be called "trustworthy AI." In fact, the terminology might end up being used as a marketing tool more than to convey factual information about the AI system. In doing so, economic appeal may usurp ethical and legal rationales.

This is particularly salient as a consideration for the review undertaken in the next section, 3.2. In particular, it helps to place this concern within real-world circumstances, demonstrating that in fact the term already seems to be used as a

marketing tool despite good intentions, and underlining why the high-level adoption of this term in international policy discourse, as discussed in Section 3.2.d., may actually backfire on the original ambition of this term.

3.2. Are you for real? It all boils down to memetic appeal

Short of finding a new term, all of the points put forward under Section 3.1. suggest that it could be (at minimum) worth changing the term when used in its original sense (i.e. to indicate adherence to all three pillars as advocated by the AI HLEG). For example, we could adopt the term "trustworthy AI™." Such a division would allow us to effectively distinguish between what the original term conceptualizes versus what people may change it to mean. "Trustworthy AI™" in its original composition appears to have had significant appeal to policymakers across the globe as Sections 3.2b-d will highlight similar terms, ideas and concepts that have populated the policy discourse in recent times.

The following sections build on the aforementioned concerns about the conflation of meanings. In particular, they build the case that the term is prone to being 'washed clear' of its original meaning. In doing so, it investigates the shift "trustworthy AI™" has undergone in international policy discourse and how the term has been (mis)appropriated and changed throughout the course of its adaptation into different contexts.

In the following sections, I establish the salience and timeliness of this discussion. First, I briefly introduce two regulatory developments to highlight the rapidly changing strategic and political landscape within which this term is being used. I will then highlight a number of relevant international policy efforts that seem to make use of versions of "trustworthy AI™" and are adjacent in intent to the EU's original ambition. This will provide a better understanding of (1) how the EU's

conceptualization may have shaped international conversation and (2) how the term has been (re)used to hold multiple meanings and was consequently watered down to a suitcase word (Minsky, 2007), void of specific meaning, content and subsequently, actionable intent.[31]

3.2.a. Political shifts: Salience and timeliness

Recent research indicates that the number of bills passed into law containing references to "AI" rose from 1 in 2016 to 18 in 2021 across 25 countries (Zhang et. al., 2022). This includes approaches to tackle ethical concerns related to AI. Over the past years, we have seen an increase in concern about the use of AI-based technology in hiring decisions due to their opaqueness and associated ethical issues such as a lack of ability to challenge the algorithm, potential bias, etc. One legal example that tackles this is the Artificial Intelligence Video Interview Act (820 ILCS 42, 2020).[32] It is the first US state law regulating the use of AI during the evaluation of prospective employees' interviews. Depending on the algorithm and the intellectual property (IP), it is often difficult or impossible to gain a full understanding of the reasoning that leads to an applicant's final outcome, rank or score. This is particularly relevant to requirement II of the Artificial Intelligence Video Interview Act, denoting that each job applicant shall be informed of how the AI-based system works and what characteristics it uses to evaluate candidates. In short, the law requires employers to: (i) notify applicants in a written format that AI may be used to analyze their video interview, (ii) give the applicants information about the workings of the AI and the characteristics it uses for evaluation of the

[31] Without actionable intent it will be difficult, if not impossible, to hold governments and industry accountable if they fail to live up to their promises. They will always be able to minimize what the term means. It should be noted that this is happening with multiple AI-related terms and not only with the term that is the focus of the paper. Other terms that come to mind are "explainability" and "accountability," which can denote quite different things depending on context and audience and are often used without any clear reference point, leading to confusion for the reader.

[32] It should be noted that the law does not define 'artificial intelligence' — creating difficulties to clearly delineate which systems the law applies to — and that the law solely addresses artificial intelligence-based technology used in videos recorded of the interview by the employer.

interview, and (iii) obtain the applicant's consent to use the aforementioned artificial intelligence. Applicants can request the deletion of their video file and employers are prohibited to share the video file beyond those actors necessary to evaluate it.

More recently, the EU has become the first governmental actor to put forward a horizontal regulatory framework for high-risk AI systems and a ban for certain AI systems with the AI Act (European Commission, 2021). In this regulatory proposal, a number of proposed legal obligations have been outlined which a high-risk AI system must fulfill through a conformity assessment before it can be deployed on the EU market.[33] An AI system is considered "high-risk" if it falls under certain categories outlined in Annex III of the AI Act (such as certain areas of access to education or employment).[34] This regulatory proposal is currently discussed in the European Parliament, the Council of the European Union and the European Commission. Amendments have already been proposed by various member states holding the presidency of the Council of the EU, ranging from tackling general purpose AI systems to real world testing. It has also received over 3000 tabled amendments in the European Parliament following the first published report (European Parliament, 2022) on the AI Act by the two lead committees on the file, the Committee on Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs.

Developments such as the aforementioned corroborate the overall sense that there is increased attention and concern about AI and its impacts on society from governments, policymakers and legislators alike. With the increased discussion, it is

[33] Interestingly, the legal obligations in the AI Act closely match the seven key requirements outlined in the Ethics Guidelines for Trustworthy AI proposed by the AI HLEG, presented earlier in this paper.

[34] Interestingly, the legal obligations in the AI Act closely match the seven key requirements outlined in the Ethics Guidelines for Trustworthy AI proposed by the AI HLEG, presented earlier in this paper.

increasingly important to ensure that concepts, terminology, methods and measures are accurate and coherent.

3.2.b. EU-adjacent efforts

Many EU member states have adopted the EU's concept for use within their own AI strategies and policy documents. These member states include Czechia (Czech Republic, 2019), Luxembourg (Luxembourg, 2019), Malta (Malta, 2019a-d), and the Netherlands (Netherlands, 2019). This is unsurprising given (a) the novelty of the concept at a time when many countries did not yet have fully fleshed-out AI strategy and (b) the fact that the EU mandated a similar and cooperative approach to AI governance to combat fragmentation, as outlined in the Declaration on Cooperation (European Commission, 2018c) and the EU's Coordinated Plan (European Commission, 2018b).

Relatedly, the recent AI Act (European Commission, 2021) indicates that standards could play a crucial role in the conformity assessment procedures of high-risk AI systems. It leaves scope for standards or technical specifications to replace matching aspects in the conformity assessment, which high-risk AI systems would need to undergo otherwise. Currently, no matching standards exist. However, in light of the overall discussion in this paper, many of the legal obligations in the AI Act's (European Commission, 2021) conformity assessment match the areas under "trustworthy AI™." This means that the original conceptualization has been highly influential beyond ethical considerations, feeding into regulatory and standardization efforts. It is noteworthy then, that standardization committees have now started working on trustworthy AI as a topic.

One group that has adopted the term at a high level is the ISO Committee IEC TR 24028:2020[35] on "Information technology — Artificial intelligence — Overview of

[35] See: https://www.iso.org/standard/77608.html.

trustworthiness in artificial intelligence." This working group predominantly focuses on areas related to the legal obligations for the conformity assessment outlined in the AI Act under Art. 11, "Annex IV Technical Documentation," and Art. 15, "Accuracy, Robustness and Cybersecurity." This matches at least one of the ethical key requirements outlined in the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) and demonstrates how the term's original conceptualization has affected concrete policy action. Moreover, it indicates that standards committees have adopted some of the terminology and thinking of the EU to, at the very least, preempt upcoming regulation such as the AI Act (European Commission, 2021).

Other standardization efforts, such as the US Senate Bill "S.1849 — Leadership in Global Tech Standards Act of 2021" (S.1849, 2021), do not make any reference to trustworthy AI.

3.2.c. International partnerships, agreements and cooperation pipelines

Many international partnerships have used the term since its inception, often diluting and modifying the original meaning.

The OECD Recommendations on AI (OECD, 2019), signed by over 35 countries, recognize that "the trustworthiness of AI systems," (p. 6) is a key factor of AI diffusion and consider it vital to foster the "adoption of trustworthy AI in society and to turning AI trustworthiness into a competitive parameter in the global marketplace," (p. 6) Not only is the adoption of the original term evident, it is, moreover, clear that the trustworthiness of AI systems is seen as a competitive advantage. This matches policy documents from the EU, such as the Communication on Building Trust in Human Centric AI (European Commission, 2019). There, the EU envisions its approach to trustworthy and human-centric AI as one that strengthens its reputation for safe, reliable and ethical products. At the

same time, although the OECD Recommendations on AI (OECD, 2019) match many of the original requirements outlined in the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019), it does not match the full scope of "trustworthy AI™."

The 2019 G20 Ministerial Statement on Trade and Digital Economy (G20, 2019) reflects the EU's vision as well. In particular, the section on the G20 AI Principles (G20, 2019) focuses on the "responsible stewardship of trustworthy AI," (p. 1) and "national policies and international co-operation for trustworthy AI," (p. 3). These Principles refer back to the OECD Recommendations on AI (OECD, 2019), and do not define trustworthy AI. As such, they have completely detached themselves from the original meaning of "trustworthy AI™," adopting what appears to be an intermediary use of the term without its original context.

The 2022 UNESCO Recommendations on the Ethics of Artificial Intelligence (UNESCO, 2022) also mention that the "the trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody the values and principles set out in this Recommendation," (p. 18).

Most recently, the term has permeated international cooperation discourse, as evidenced by the EU-US Trade and Technology Council's inaugural Pittsburgh statement (EU-US Trade and Technology Council, 2021). This states that "The European Union and the United States affirm their willingness and intention to develop and implement trustworthy AI and their commitment to a human-centered approach that reinforces shared democratic values and respects universal human rights, which they have already demonstrated by endorsing the OECD Recommendation on AI." (p.11). It is interesting that despite the EU being one of the two main leads in this discourse, the document itself refers back to the OECD

Recommendations on AI (OECD, 2019) and not to the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) when referring to trustworthy AI.

3.2.d. International actors

Beyond international fora, AI strategies of international powers such as the United States have equally explored versions of trustworthy AI. There appears to have been a distinct uptake in the use of the term in US policy documents subsequent to the publication of the draft version of the final Ethics Guidelines for Trustworthy AI (AI HLEG, 2018).[36]

A preliminary review compared and contrasted documents published before and after the *draft* Ethics Guidelines for Trustworthy AI (AI HLEG, 2018), which first mentioned the concept of trustworthy AI. The documents covered were the following: Preparing for the Future of Artificial Intelligence (NSTC, 2016a); The National Artificial Intelligence Research and Development Strategic Plan (NSTC, 2016b); Artificial Intelligence, Automation, and the Economy (Executive Office of the President, 2016); The FUTURE of Artificial Intelligence Act of 2017 (US Senate, 2017); the Algorithmic Accountability Act of 2019 (H.R.2231, 2019); Supporting the development of guidelines for ethical development of Artificial Intelligence (HRES 153, 2019); notes from the Office of Science and Technology's Select Committee on Artificial Intelligence's inaugural meeting June 27, 2018; AI Principles;[37] Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense (DIB, 2019); National Security Strategy of the United States of America (White House, 2017); Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and

[36] However, given the number of similar sounding terminologies that have been in use, including aspects discussed in Section 1 of this paper, it is difficult to say with certainty whether the US policy space's frame of reference has indeed been shaped by the EU or whether they decided to use this term independently.

[37] See: https://epic.org/wp-content/uploads/privacy/ai/WH-AI-Select-Committee-First-Meeting.pdf.

Prosperity (DoD, 2018), and the request for comments on a Draft Memorandum to the Heads of Executive Departments and Agencies on the subject of 'Guidance for Regulation of Artificial Intelligence Applications'.[38] It should be noted that none of these documents published prior to the draft Ethics Guidelines for Trustworthy AI (AI HLEG, 2018) mention a variation of 'trustworthy AI', this includes two reviewed documents put forward after the publication of the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019).

In publications under the Trump administration, trustworthy AI as a concept notably begins featuring in the 2019 Interim report of the National Security Commission on Artificial Intelligence (NSCAI, 2019); the Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (E.O. 13960, 2020); the National Artificial Intelligence Research and Development Strategic Plan: 2019 Update (NITRD, 2019); and in the NIST's U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (NIST, 2019). The Executive Order on Maintaining American Leadership in Artificial Intelligence (E.O. 13859, 2019), which was published prior to the final Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) but subsequent to their draft version (AI HLEG, 2018) does not use the term "trustworthy AI." Yet, it comes close by referencing the "development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies" (p. 3970).

Similar to other documents introduced under Section 3.2. many international policy documents ended up using "trustworthy AI" without properly defining it or referencing the EU's definition. For example, the Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (E.O. 13960, 2020) has "trustworthy AI" in its title, but lacks concrete conceptualization throughout the the text. On the other hand, the NIST's U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (NIST, 2019) mentions "reliable, robust, and trustworthy AI technology

[38] See: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

development" (p. 4, p. 22). While it never concretizes the term, "reliability" and "robustness" match a subset of the original "trustworthy AI's™" three pillars (i.e. pillar three).

Looking at AI governance efforts in China, according to CSET's translation of the 2021 Ethical Norms for New Generation Artificial Intelligence[39] (original text source: MOST, 2021) published by the PRC's Ministry of Science and Technology, this policy document references trustworthiness at large. In particular, Art. IV of the document outlines norms for "Assurance of Controllability and Trustworthiness." However, it does not use the term "trustworthy AI" or refer to the EU's efforts in that space.

3.3. Lessons

There appear to be two intertwined lessons we can draw from the preceding sections. First, given the rapid rise in AI strategies, there was a **significant policy vacuum** that raised new considerations. This contributed to a situation in which governments and policymakers were under significant time pressure to develop relevant discourse. This, in turn, made it more appealing to align work with that of others who faced similar issues and to handle them in a similar manner, be that in concept or content. One area where this 'copy-paste' discourse was especially evident was in the development of Ethics Guidelines. Because all actors arrived at similar conclusions and were inconspicuously inspired by similar texts, almost all Ethics Guidelines ended up developing a similar subset of areas (Hagendorff, 2020; Jobin, Ienca and Vayena, 2019).[40] Generally speaking, it makes sense that well thought-out and researched ideas would proliferate similar discourse and inspire adjacent strategies and documents. In many cases, this may even be a good thing,

[39] See: https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.
[40] It should be noted, however, that it is difficult to capture an entire field and that this coherence may also be due to a simplification and accessibility concern for audiences.

particularly if the original concept or idea is robust and translated into many different documents by virtue of reference or adaptation, without it being watered down or amended. A similar process applied to the conceptualization of what overarching class of AI systems governments wish to encourage, and thus "trustworthy AI™" inspired a class of efforts to use a variation of the term in their own documents and processes, for better or worse.

Second, the preceding sections demonstrate that the EU has an **opportunity to shape the AI policy space**. Section 3 supports the idea that the EU has been wielding some degree of soft power with "trustworthy AI™" (albeit only in description and not in content). This may be indicative of the vacuum with regards to good AI policy approaches, as described above, as well as of the memetic appeal of the original term.

The belief that the EU has some capability of influencing non-EU actors to adopt its laws and policies is often referred to as the "Brussels effect" (coined by Anu Bradford (Bradford, 2020) and similar to the California effect[41] in the US).[42] Extrapolating from the memetic effect the term "trustworthy AI" has had, it will be interesting to see the degree to which policy proposals or regulatory proposals — such as the AI Act (European Commission, 2021) — will shape international policy discourse and development. For example, with regard to the AI Act (European Commission, 2021), there is already some concrete evidence that supports the Brussels effect. In particular, the Canadian government recently published their Artificial Intelligence and Data Act (Bill C-27, 2022), a draft act to regulate AI. This draft act clearly orients itself on the AI Act (European Commission, 2021) by taking a risk-based approach to regulating AI and proposing requirements that a select group of AI systems would need to adhere to (described as "high impact", in the AI

[41] See: https://en.wikipedia.org/wiki/California_effect.
[42] One of the most commonly cited examples of the Brussels Effect is the European General Data Protection Regulation which California emulated when it passed the CCPA: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

Act these would be described as “high-risk”). Most recently, the Brussels effect has been explored in scholarly discourse evaluating how the EU may shape strategic regulatory interventions internationally against a specific set of criteria (Siegmann & Anderljung, 2022).

It should be noted for completeness that the AI Act itself builds on previous formulations of risk-based governance approaches, such as the German national AI strategy (German Federal Government, 2020; Lütge et al., 2022).

4. Conclusion

This paper provides the background of the development, conceptualization, and proliferation of the term “trustworthy AI,” as advanced by the EU in government discourse. It elaborates on similar terminologies that arose during the same time frame and puts forward a select number of concerns with regard to the term. In reviewing various international efforts, both collaborative and individual, this paper illustrates that the term has had broad appeal to policymakers across and outside of the EU. While it is too early to tell, it is likely that the memetic impact of this term, and the EU’s first mover advantage in defining it, will be replicated in the EU’s more consequential policy and regulatory efforts, most notably the AI Act (European Commission, 2021), suggesting a possible Brussels effect.

Bibliography

AI HLEG. 2018. “Draft Ethics Guidelines for Trustworthy Artificial Intelligence.” Independent High-Level Expert Group on Artificial Intelligence. https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEGDraftAIEthicsGuidelinespdf.pdf

AI HLEG. 2019. “Ethics Guidelines for Trustworthy Artificial Intelligence.” Independent High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

ALLAI. 2021. “EU Proposal for Artificial Intelligence Act, Analysis and Recommendation.” .https://allai.nl/wp-content/uploads/2021/08/EU-Proposal-for-Artificial-Intelligence-Act-Analysis-and-Recommendations.pdf

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. “Concrete Problems in AI Safety.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1606.06565.

Anderson, James M., Nidhi Kalra, Karlyn Stanley, Paul Sorensen, Constantine Samaras, and Tobi A. Oluwatola. 2016. “Autonomous Vehicle Technology: A Guide for Policymakers.” RAND Corporation. https://www.rand.org/pubs/research_reports/RR443-2.html.

Barocas, Solon, and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” *An International Journal on Information Fusion* 58 (June): 82–115.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. “Fairness in Criminal Justice Risk Assessments: The State Of the Art.” *Sociological Methods & Research*, July, 004912411878253.

Bietti, Elettra. 2021. “From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics.” *Journal of Social Computing 2*(3), 266–83.

Bill C-27. 2022. “An Act to enact the Consumer Privacy Protection Act, the Personal

Information and Data Protection Tribunal Act and the *Artificial Intelligence and Data Act* and to make consequential and related amendments to other Acts." House of Commons of Canada. Minister of Innovation, Science and Industry. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading.

Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al., 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2004.07213.

Bryson, J. 2018. "No One Should Trust Artificial Intelligence." *Science & Technology: Innovation, Governance* 11, 14.

Calo, Ryan. 2010. "Peeping HALs: Making Sense of Artificial Intelligence and Privacy." *EJLS - European Journal of Legal Studies* 2(3), 168–92.

Christian, Brian. 2021. *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books.

Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.

Coeckelbergh, Mark. 2020. *AI Ethics*. MIT Press.

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. 2019. "AI Now 2019 Report." AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.pdf.

Czech Republic. 2019. "National Artificial Intelligence Strategy of the Czech Republic." Ministry of Industry and Trade. https://www.mpo.cz/assets/en/guidepost/for-the-media/press-releases/2019/5/NAIS_eng_web.pdf.

Dafoe, Allen. 2020. "AI Governance: Opportunity and Theory of Impact." *Effective Altruism Forum,* Accessed 17 September, 2022.

Danaher, John. 2019. "Automation and Utopia: Human Flourishing in a World without Work." Harvard University Press.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. (2022). "The AI Index 2022 Annual Report." AI Index Steering Committee. Stanford University. Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf.

DIB. 2019. "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense". Defense Innovation Board. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

Dignum, Virginia. 2019. "Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way." Springer Nature.

DoD. 2018. "SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY: Harnessing AI to Advance Our Security and Prosperity". United States Department of Defense. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

Doran, Derek, Sarah Schulz, and Tarek R. Besold. 2017. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives." *arXiv:1710.00794 [cs]*, October. http://arxiv.org/abs/1710.00794.

European Commission. 2018a. "Communication from the Commission - Artificial Intelligence for Europe (COM(2018) 237 final)." Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

European Commission. 2018b. "Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Coordinated Plan on Artificial Intelligence (COM/2018/795 final)." Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN.

European Commission. 2018c. "(Digital Day) Declaration on Cooperation on Artificial Intelligence." European Commission website - JRC Science Hub - Communities. https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/eu-d

eclarat ion-cooperation-artificial-intelligence.

European Commission. 2019. “Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human-Centric Artificial Intelligence (COM/2019/168 final).” Brussels: European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168.

European Commission. 2021. “Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM/2021/206 final).” Brussels: European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-1 1eb-958 5-01aa75ed71a1.

European Parliament. 2022. “Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD)).” Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs. https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563_EN.pdf.

EU-US Trade and Technology Council. 2021. “Pittsburgh Statement.” Joint Declaration. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_21_4951.

Executive Office of the President. 2016. “Artificial Intelligence, Automation, and the Economy.” https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF.

Executive Office of the President. 2019. E.O. 13859. “Executive Order on Maintaining American Leadership in Artificial Intelligence.” https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.

Executive Office of the President. 2020. E.O. 13960. “Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.” https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

Fischer, Sophie-Charlotte, Jade Leung, Markus Anderljung, Cullen O’keefe, Stefan

Torges, Saif M. Khan, Ben Garfinkel, et al. 2021. “AI Policy Levers: A Review of the U.s. Government’s Tools to Shape AI Research, Development, and Deployment.” Accessed June 1, 2022. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” *Technological Forecasting and Social Change* 114: 254–80.

Gasser, Urs. 2016. “Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy.” 2016. https://harvardlawreview.org/2016/12/recoding-privacy-law-reflections-on-the-future-relationship-among-law-technology-and-privacy/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.

German Federal Government. 2020. “Artificial Intelligence Strategy of the German Federal Government.”
https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf

G20. 2019. “Ministerial Statement on Trade and Digital Economy.” https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf.

Hagendorff, Thilo. 2020. “The Ethics of Ai Ethics: An Evaluation of Guidelines.” *Minds and Machines*, 1–22.

Hagerty, Alexa, and Igor Rubinov. 2019. “Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1907.07892.

H.Res.153. 2019. “Supporting the development of guidelines for ethical development of artificial intelligence.”

House resolution 153. https://www.congress.gov/bill/116th-congress/house-resolution/153/text.

H.R.2231. 2019. “Algorithmic Accountability Act of 2019”. House Bill 2231. https://www.congress.gov/bill/116th-congress/house-bill/2231/text.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “Artificial Intelligence: The Global Landscape of Ethics Guidelines.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1906.11668.

Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. “AI Safety Gridworlds.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1711.09883.

Lockey, Steven, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. “A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions.” In *Hawaii International Conference on System Sciences 2021 (HICSS-54)*. https://aisel.aisnet.org/hicss-54/os/trust/2/.

Luxembourg. 2019. “Artificial Intelligence: a strategic vision for Luxembourg.” The Government of the Grand Duchy of Luxembourg. https://digital-luxembourg.public.lu/sites/default/files/2019-05/AI_EN.pdf.

Lütge, Christoph, Hohma, E., Boch, A., Poszler, F. & Corrigan, C. 2022. “White Paper – On a Risk-Based Assessment Approach to AI Ethics Governance.” IEAI, https://www.ieai.sot.tum.de/wp-content/uploads/2022/06/IEAI-White-Paper-on-Risk-Management-Approach_2022-FINAL.pdf.

Maas, Matthijs M. 2021. “Aligning AI Regulation to Sociotechnical Change.” https://doi.org/10.2139/ssrn.3871635.

Malta. 2019a. “Malta the ultimate AI Launchpad: a strategy and vision for Artificial Intelligence in Malta 2030.” https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.

Malta. 2019b. “Malta towards an AI strategy: High-level document for public consultation.” https://malta.ai/wp-content/uploads/2019/04/Draft_Policy_document_-_online_version.pdf.

Malta. 2019c. “MCAST: Artificial Intelligence Strategy: roadmap 2025.” Malta College of Arts, Science & Technology. https://www.mcast.edu.mt/wp-content/uploads/AI-Strategy_Final.pdf.

Malta. 2019d. “Malta towards trustworthy AI: Malta’s Ethical AI Framework.” https://malta.ai/wp-content/uploads/2019/10/Malta_Towards_Ethical_and_Trustworthy_AI_vFINAL.pdf.

Minsky, Marvin. 2007. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon and Schuster.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. 2021. "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics." *Minds and Machines* 31 (2): 239–56.

Morley, Jessica, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2021. "Operationalising AI Ethics: Barriers, Enablers and Next Steps." *AI & Society*, November. https://doi.org/10.1007/s00146-021-01308-8.

MOST. 2019. "Ethical Norms for New Generation Artificial Intelligence". The National New Generation Artificial Intelligence Governance Specialist Committee (国家新一代人工智能治理专业委员会). PRC Ministry of Science and Technology (MOST; 科学技术部; 科技部). http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

Mökander, J., Juneja, P., Watson, D.S. et al. (2022). "The US Algorithmic Accountability Act of 2022 vs.The EU Artificial Intelligence Act: what can they learn from each other?". Minds & Machines. https://doi.org/10.1007/s11023-022-09612-y

Müller. 2021. "Ethics of Artificial Intelligence 1." *The Routledge Social Science Handbook of AI*. https://doi.org/10.4324/9780429198533-9/ethics-artificial-intelligence-1-vincent-mã¼ller.

Netherlands. 2019. "Strategisch Actieplan voor Artificiële Intelligentie." Ministerie van Economische Zaken en Klimaat. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voorartificiele-intelligentie/Rapport+SAPAI.pdf.

NIST. 2019. National Institute of Standards and Technology, US Department of Commerce. "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools." https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

NITRD. 2019. "National Artificial Intelligence Research and Development Strategic Plan: 2019 Update." A Report by the Select Committee on Artificial intelligence of the National Science and Technology Council. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

NSCAI, 2019. “Interim report of the National Security Commission on Artificial Intelligence, National Security Commission on Artificial Intelligence.” https://epic.org/wp-content/uploads/foia/epic-v-ai-commission/AI-Commission-Interim-Report-Nov-2019.pdf.

NSTC. 2016a. “Preparing for the Future of Artificial Intelligence.” Executive Office of the President. National Science and Technology Council. Subcommittee on Machine Learning and Artificial Intelligence. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

NSTC. 2016b. “The National Artificial Intelligence Research and Development Strategic Plan”. Executive Office of the President. National Science and Technology Council. Networking and Information Technology Research and Development Committee. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.

Nyholm, Sven, and Jilles Smids. 2016. “The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?” *Ethical Theory and Moral Practice: An International Forum* 19 (5): 1275–89.

OECD. 2019. “Recommendation of the Council on Artificial Intelligence.” C(2019)34, C/MIN(2019)3/FINAL, OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

ÓhÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. 2020. “Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance.” *Philosophy & Technology* 33 (4): 571–93.

Ramus, Catherine A., and Ivan Montiel. 2005. “When Are Corporate Environmental Policies a Form of Greenwashing?” *Business & Society* 44 (4): 377–414.

Russel, S. and Norvig, P. (1995). “Artificial Intelligence: A Modern Approach.” *Prentice Hall Upper Saddle River,* NJ. https://www.sti-innsbruck.at/sites/default/files/Knowledge-Representation-Search-and-Rules/Russel-&-Norvig-Inference-and-Logic-Sections-7.pdf.

Ryan, Mark, and Bernd Carsten Stahl. 2020. “Artificial Intelligence Ethics Guidelines for Developers and Users: Clarifying Their Content and Normative Implications.” *Journal of Information, Communication and Ethics in Society* 19 (1): 61–86.

Salo-Pöntinen, H., & Saariluoma, P. 2022. “Reflections on the human role in AI

policy formulations: how do national AI strategies view people?" Discover Artificial Intelligence, 2, Article 3. https://doi.org/10.1007/s44163-022-00019-3

Samoili, Sofia, Montserrat Lopez Cobo, Emilia Gomez, Giuditta De Prato, Fernando Martinez-Plumed, and Blagoj Delipetrev. 2020. "AI Watch. Defining Artificial Intelligence. Towards an Operational Definition and Taxonomy of Artificial Intelligence." https://eprints.ugd.edu.mk/28047/.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58. AIES '20. New York, NY, USA: Association for Computing Machinery.

SDG. 2015. "Sustainable Development Goals." https://sdgs.un.org/goals

Siegmann, C., Anderljung, M. (2022). The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market. https://uploads-ssl.webflow.com/614b70a71b9f71c9c240c7a7/62fbe1c37eff7d304f0803ac_Brussels_Effect_GovAI.pdf

Stix, Charlotte. 2021a. "Actionable Principles for Artificial Intelligence Policy: Three Pathways." *Science and Engineering Ethics* 27 (1): 15.

———. 2021b. "The Ghost of AI Governance Past, Present and Future: AI Governance in the European Union." *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2107.14099.

UNESCO. 2022. "Recommendations on the Ethics of Artificial Intelligence." adopted November 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

U.S. Senate. 2017. "The FUTURE of Artificial Intelligence Act of 2017. BAG17H16. https://www.cantwell.senate.gov/imo/media/doc/The%20FUTURE%20of%20AI%20Act%20Introduction%20Text.pdf.

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals." *Nature Communications* 11 (1): 1–10.

Wang, P. 2019. "On Defining Artificial Intelligence." *Journal of Artificial General Intelligence*. https://sciendo.com/downloadpdf/journals/jagi/10/2/article-p1.pdf.

Winter, Philip Matthias, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler.

2021. “Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/2103.16910.

White House. 2017. “National Security Strategy of the United States of America.” https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf.

117th Congress. 2021-2022. “S.1849 - Leadership in Global Tech Standards Act of 2021.” https://www.congress.gov/bill/117th-congress/senate-bill/1849.

820 ILCS 42/. 2020. “Artificial Intelligence Video Interview Act.” https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68.

Chapter II

The ghosts of AI governance past, present and future: AI governance in the European Union

……..

In: Justin Bullock & Valerie Hudson (eds.), *Oxford University Press Handbook on AI Governance, Section 9: International Politics and AI Governance*
(Oxford University Press, forthcoming).

……..

Introduction

This chapter will provide a simplified overview of the past, present and future of AI governance in the EU. It will give the reader a solid background understanding of how the EU reached the current status as global leader in the regulation of AI, how all the different pieces are interconnected and where the EU might go next. Section 1 will discuss a select number of EU policy efforts of the past years and illustrate how they built on each other. It is argued that, by virtue of spearheading 'trustworthy AI', the EU has occupied a position where it has been able to shape the discourse on global AI governance discourse early on. Section 2 will introduce the history of the EU's AI regulation, proposed in April 2021, highlighting its connection to previous policy efforts, and its roots in adjacent measures to strengthen the EU's technological ecosystem. Finally, in Section 3, the author will turn to the future, considering and exploring a number of AI governance areas that are prime candidates to become crucial for AI governance in the EU in the coming decade.

1. The Past

Taking stock: the roads towards the EU's AI governance

This section will highlight the most relevant and recent EU policy developments with regard to AI, illustrate how they contributed to shaping both the broader narrative for the EU and how they set the cornerstones for AI governance in the EU. It will be suggested that the AI Act (European Commission, 2021c), the European Commission proposal for a horizontal AI regulation and the accompanying policy measures in the EU form a coherent and strategically aligned link in a chain of policies which were initiated many years ago. To that end, some of these policy documents will be revisited in Section 2, which presents the different elements that form the bigger picture of current AI governance in the EU.

While going through the formative policy developments, it is worth underlining that the European Commission has invested in and funded AI and AI-related research and innovation projects for much longer than they have been focusing on the governance of AI, notably, under Horizon 2020 and before. With that in mind, the following paragraphs will set out the main elements that led to the EU to push for ethical governance of AI and to make it their guiding principle for accompanying policy measures.

1.1. The roads that led us here

The earliest key policy document is the resolution by the European Parliament with 'Recommendations to the Commission on Civil Law Rules on Robotics' in 2017 (henceforth: 'Civil Law Rules on Robotics', European Parliament, 2017). Although not yet referring to AI directly in the title, the resolution laid one of the first cornerstones for the succeeding process from the European Parliament's side, by suggesting that the EU's legal framework should be updated and complemented by ethical principles on the topic, that the environmental impact of AI and robotics

should be kept low, and that the societal and economic impacts of future systems deserve heightened attention.[43]

Shortly thereafter, the European Economic and Social Committee (EESC) presented their 'Opinion on AI' (European Economic and Social Committee, 2017). The 'Opinion on AI' (European Economic and Social Committee, 2017) discusses the need to verify, validate and monitor AI, as well as AI-based systems, advocates for an overarching "human-in-command" approach and leans into the necessity for ethical, societal and safety considerations. Accordingly, recommendations cover the development of a code for ethics, a ban on Lethal Autonomous Weapons Systems and the development of suitable standardization systems for AI.

On evaluation, we can already see that while they are among the earliest EU policy documents on the topic, the EP's 'Civil Law Rules on Robotics' (European Parliament, 2017), and the EESC's 'Opinion on AI' (European Economic and Social Committee, 2017) have some overlaps. These include a demand for inclusion of the ethical dimension in the discussion and an acknowledgement of the societal impact, alongside proposals for recourse.

The shift towards EU AI governance as a topic largely independent of robotics from the perspective of EU institutions, was further solidified through the European Group on Ethics in Science and New Technologies (EGE) Statement on 'Artificial Intelligence, Robotics and Autonomous Systems' (henceforth: Statement; European Group on Ethics in Science and New Technologies, 2018). An independent advisory body to the European Commission, the EGE advises it on the intersection of science and emerging technologies with ethical, societal and fundamental rights issues. In their 'Statement' (European Group on Ethics in Science and New Technologies, 2018), they echoed the need to establish an overarching framework on AI in the EU with an ethical dimension. The goal would be to tackle the ethical, legal and societal

[43] We will revisit the role this Report continues to play in Section 3.

governance issues, ensuring that AI is created with "humans in mind" (European Group on Ethics in Science and New Technologies, 2018). Therefore, in a sense building on the resolution on 'Civil Law Rules on Robotics' (European Parliament, 2017) and 'Opinion on AI' (European Economic and Social Committee, 2017), they proposed the development of several ethical Principles for AI based on fundamental European values. This proposal was both quintessentially European, outlining the importance of fundamental rights and values, and well timed to fit within the broader international landscape, where principles for AI were starting to see their advent (Fjeld et al., 2020; Hagendorff, 2019; Schiff et al., 2020; Zeng et al., 2018).

These three documents together could be seen as the first heralds of where the EU is now: regulating AI with a focus on human-centricity and ethics. They demonstrated what aspects of AI were considered important by the EU's legislative body, the EU's civil society organization body and the main EU group on ethics at that point. This is a convergence point where demonstrable attention has begun from a legislative angle, a societal angle and an ethical angle. As we will see in Section 2 this interplay has since continued, and even been strengthened. At this point in time it became strikingly evident that the EU policy space was alert to the challenges posed by AI, as much as to the opportunities AI could hold, and the necessity for a more methodological approach became pressing.

What followed were major leaps. First, the European Commission presented the 'Digital Day Declaration on Cooperation on AI' (European Commission, 2018c) in April 2018. Second, they responded to the call from the European Council to "put forward a human-centric approach to AI"[44] by presenting their AI strategy in the Communication entitled 'Artificial Intelligence for Europe' (European Commission, 2018a), in the same month. I propose that these documents foreshadow current EU AI governance mechanisms. The 'Digital Day Declaration on Cooperation on AI' (henceforth: 'Declaration'; European Commission, 2018c) anticipated the

[44] See: https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf.

'Coordinated Plan on AI' (European Commission, 2018b), whereas the 'Communication on AI for Europe' (European Commission, 2018a), in some sense preceded the proposal for a regulatory framework, the AI Act (European Commission, 2021c).

So, *how did these documents set out the European Commission approach to AI governance?*

In the 'Declaration' (European Commission, 2018c), signed by all 28 Member States in 2018 (at that time including the United Kingdom) as well as Norway, the countries agreed to engage in close dialogue with the European Commission on the topic of AI and coordinate their actions. I propose that this is the first instance internationally where a significant number of countries agreed to coordinate on AI governance.[45] International, later-stage efforts to coordinate among multiple countries such as the Global Partnership on AI (GPAI) or the OECD included the EU by way of representation through the European Commission, as well as a subset of member states. As of this writing, there is no other international AI governance effort that envisages the same level of coordination, alignment of approach, and pooling of resources as the one started with the 'Declaration'[46] and signed off by the member states.[47]

We have seen that the 'Opinion on AI' (European Economic and Social Committee, 2017) called for a "human in command" approach and that the EGE 'Statement' (European Group on Ethics in Science and New Technologies, 2018) called for AI to be made with "humans in mind". This notion of human-centric AI is revisited in the 'Declaration' (European Commission, 2018c), committing signatories to ensure that 'humans remain at the center of AI development', and to prevent the "harmful

[45] Of course, such coordination may be seen as implicit by virtue of these countries being Member States of the European Union.
[46] This Declaration has been expanded on significantly in the Coordinated Plan on AI and its follow up.
[47] It should be noted that the Declaration is non-binding. However, any other international efforts are equally non-binding at the time of this writing.

creation and use of AI applications."

Echoing topics outlined in the resolution on 'Civil Law Rules of Robotics' (European Parliament, 2017) and the 'Opinion on AI' (European Economic and Social Committee, 2017), the 'Declaration' (European Commission, 2018c) focuses on the development of a collaborative framework to coordinate on relevant areas such as sustainability, labor market, funding and ethics. Alongside the necessary mitigation of ethical risks, it also touches upon legal and socio-economic risks of AI. All of this is underlined with the recognition that the existing ecosystem needs to be boosted in order for the EU to remain competitive and agile for future challenges.

The 'Communication on AI for Europe' (henceforward: AI Strategy; European Commission, 2018a) picked up on this and presented the EU's three-pronged strategy for AI taking into account the ecosystem, civil society as well as ethics and regulation. In short, it recommended to "(1) boost Europe's technological and industrial capacity; (2) to prepare Europe for the socio-economic changes associated with AI; and (3) to ensure that Europe has an appropriate ethical and legal framework to deal with AI development and deployment" (European Commission, 2018a).

The AI strategy (European Commission, 2018a) also outlined many ambitions that are currently relevant, such as the development of regulatory sandboxes (eventually key for a horizontal EU AI regulation) or a commitment to support centers for data sharing (for example, the EU Data Hubs).

In order to tackle the third pillar of its AI strategy (European Commission, 2018a), the European Commission set up an independent High-Level Expert Group on

Artificial Intelligence (AI HLEG) tasked with, amongst other deliverables, the development of Ethics Guidelines.[48]

Moreover, the AI strategy (European Commission, 2018a) also served to present, for the first time, a clear 'European way for AI' vis-a-vis the international stage. It clearly outlined the role that the EU envisages for itself when it comes to AI governance. It stated that;

> "the EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation and respects Union's values and fundamental rights as well as ethical principles such as accountability and transparency. The EU is also well placed to lead this debate on the global stage. This is how the EU can make a difference - and be the champion of an approach to AI that benefits people and society as a whole."

The EU clearly positioned itself as an actor on the international stage who will put ethical considerations and fundamental rights at the core of AI governance.

Finally, we end this section by pulling some of the threads together while leaving space to investigate the regulatory efforts in Section 2. Many of the efforts highlighted culminate or find resonance in the European Commission's 'Coordinated Plan on the Development and Use of Artificial Intelligence Made in Europe' (henceforth: Coordinated Plan; European Commission, 2018b) published in late 2018.

The 'Coordinated Plan' (European Commission, 2018b) picks up where the 'Declaration' (European Commission, 2018c) left off. It too was agreed on by all

[48] These will be further explored in Subsection 1.2. Another deliverable not discussed in this chapter are the Policy Recommendations for Trustworthy AI.

Member States as well as Norway and Switzerland and is to be updated on a rolling basis. It echoes plans mapped out in the AI strategy (European Commission, 2018a), namely that a European approach to AI should be built upon ethical and societal values derived from the Charter of Fundamental Rights. Moreover, going above and beyond the perspective of previous policy documents, it puts a strong emphasis on what should become the European northstar for AI, by highlighting what it perceives to be interconnected concepts of "trusted AI" and "human-centric AI" (European Commission, 2018b).

The 'Coordinated Plan' (European Commission, 2018b) paints a picture of how the Member States can coordinate their AI strategies, define a common vision and encourage synergies between ongoing efforts in the Member States – with an eye to increasing the EU's global competitiveness and to counteracting fragmentation and competition between like-minded actors. The preliminary framework for coordination homes in on a couple of focus areas such as on commonly shared societal challenges, increased diffusion of AI, support to AI excellence, data availability and a regulatory framework. The last aspect is described as a 'seamless regulatory environment' in other parts of the text, and we will see in Section 2 what the EU has developed on that front. The 'Coordinated Plan on AI' (European Commission, 2018b) also contained a commitment on the EU's side to invest EUR 20bn into AI by 2020, and scale up towards yearly investments of that sum from then until 2027.[49]

In an adjacent stream, in early 2019, the European Parliament's Committee on Industry, Research and Energy (ITRE) had their report on 'A comprehensive European industrial policy on artificial intelligence and robotics' (European Parliament, 2019) adopted, which articulated a clear need for a "robust legal and ethical framework for AI," and which amongst other things called for ethical

[49] This includes investment on Member State-level, as well as public-private partnerships, the Digital Europe Programme and Horizon Europe funding (the latter two both run between 2021-2027).

principles that are in compliance with relevant EU and national law. Along these lines it also welcomes the work of the AI HLEG, further outlined in Section 1.2. Furthermore, it proposes aspects related to e.g. personal data and privacy, consumer protection, transparency, explainability and bias. In tandem with the EU's approach (as mapped so far), the industrial policy stresses the importance of human-centric technology and to encourage ethical values with regards to AI development and deployment. Indeed, this may set the EU apart and propel it to take lead on an international stage.

Combining the various policy efforts and taking a bird's-eye view, clear directions are emerging that concern the role the EU has set for itself generally and on the international stage. Ethical concerns, fundamental rights and values play a vital role in the EU's AI governance future. To account for this, the next subsection focuses on the development of ethical principles for AI in the EU – and their subsequent impact.

The topic covered in the next subsection should be seen as an adjacent stream of work that resulted out of the landscape built by the policy initiatives outlined here which became a policy effort in its own right, eventually feeding back into the current landscape mapped in Section 2.

1.2. Coining "trustworthy AI"

This section explores how the EU came to adopt and pioneer the term "trustworthy AI" from its ethical investigations, and what this shift marked.

Subsequent to its ambition to develop an appropriate ethical and legal framework for AI, the European Commission set out to establish two groups to support the AI strategy described in its Communication on AI: the High-Level Expert Group on AI (AI HLEG) and the European AI Alliance. The latter was set up as an accessible

online multi-stakeholder platform with the goal of contributing to the work of the European Commission and the AI HLEG.

The AI HLEG, on the other hand, was set up as an independent expert group by the European Commission, populated through a selection process (Stix, 2021). The AI HLEG was tasked with the primary goal of developing ethics guidelines. The result of their work, especially the 'Ethics Guidelines for Trustworthy AI' (AI HLEG, 2019) and the 'Assessment List for Trustworthy AI' (henceforth: Assessment List; AI HLEG, 2020), were core to the recent model of AI governance in the EU as the following paragraphs will demonstrate. To that end, the focal point for the following paragraphs will be on the 'Ethics Guidelines for Trustworthy AI' (A HLEG, 2019) and the associated 'Assessment List for Trustworthy AI: for Self-Assessment' (AI HLEG, 2020), and how the conceptualization of ethical AI contained in them contributed to Europe's vision of 'trustworthy AI'.

Tasked with the development of ethics guidelines, the AI HLEG, composed of 52 experts representing various sectors and types of expertise, underwent a comprehensive process to take a unique step towards a framework for the ethical governance of AI. Although the work was conducted internally, the AI HLEG did share their progress in meetings open to institutional observers and solicited feedback on their first draft version of the Ethics Guidelines half a year into the process via the AI Alliance.[50] Following the implementation of this public feedback, the AI HLEG presented their final 'Ethics Guidelines for Trustworthy AI' (henceforth: Ethics Guidelines; AI HLEG, 2019) in April 2019.

The 'Ethics Guidelines' (AI HLEG, 2019) constituted the first document that proposed a clear conceptual understanding and framing of what type of AI should be encouraged within the EU. While the document is strongly anchored in EU values

[50] A feedback mechanism was also used in the case of the Assessment List, for which feedback was received through two online questionnaires and through in-depth interviews with different types of organizations.

and fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union, the core concept is that of 'trustworthy AI'. In this reading, 'Trustworthy AI' is to fulfill three conditions: (i) it should be lawful, (ii) it should be ethical and (iii) it should be robust (both from a technical and social perspective).

While the 'lawful' aspect is left to existing regulation and future regulatory efforts, the document proceeds to outline the other components. In particular, our focus will be on the ethical component. The AI HLEG distilled a number of core values, which informed four principles. These are; *Respect for Human Autonomy*, *Prevention of Harm*, *Fairness* and *Explicability*. From these four ethical principles, they derived their seven key requirements to achieve 'trustworthy AI' and to operationalize these four identified principles.

The seven key requirements covered:

- *Human Agency and Oversight,* which relates to the principle of *Human Autonomy* and requires that AI system's allow for human oversight, support the user's agency and foster fundamental rights.
- *Technical Robustness and Safety,* which relates to the principle of *Prevention of Harm*. It addresses concerns such as resilience to attack (e.g. through data poisoning or model leakage), the need for suitable fallback plans, reliability and reproducibility.
- *Privacy and Data Governance,* which links to the principle of *Prevention of Harm*. It tackles the initial stages of data collection (e.g. regarding the quality and integrity of the data), as much as the need for data protocols to govern data access, and overall privacy measures throughout the AI life cycle.
- *Transparency,* which links to the principle of *Explicability*. This means, among other things, that traceability should be ensured and that capabilities and intentions (both from a technical POV and from an industry perspective) should be clearly communicated.

- *Diversity, Non-Discrimination and Fairness,* which links to the principle of *Fairness*. It states that all affected stakeholders throughout the AI life cycle need to be taken into consideration and duly involved. This means e.g. ensuring equal access and equal treatment.
- *Societal and Environmental Well-Being,* which relates to both the principle of *Fairness* and the principle of *Prevention of Harm*. It relates to the broadest range of stakeholders, the environment and the wider society. Considerations are e.g. AI's social impact and the sustainability of the current AI supply chain.
- *Accountability,* which is the last key requirement and ties all the previous requirements together. It is informed by the principle of *Fairness*. It focuses on redress mechanisms, trade-offs between principles and the need to have adequate mechanisms in place to report potential negative impacts.

At this point, the threads started with the report on 'Civil Law Rules on Robotics' (European Parliament, 2017), the 'Opinion on AI' (European Economic and Social Committee, 2017) and the AI Strategy (European Commission, 2018a) have come to reach a fuller picture: the EU's ambition to create an ethical approach towards the development and deployment of AI and an appropriate ethical framework has become a reality.

In order to operationalize these key requirements further, the 'Ethics Guidelines' (AI HLEG, 2019) also contained a draft Assessment List which was piloted and revised in the second year of the group's mandate.[51] The final 'Assessment List for Trustworthy AI: for Self-Assessment' (henceforth: 'Assessment List'; AI HLEG,

[51] The European Commission opened a broad stakeholder consultation process where feedback was solicited through three different streams: (a) a quantitative stream, (b) a qualitative stream and (c) a holistic stream. The quantitative stream consisted of two surveys, one for developers and deployers, and one for other stakeholders. The qualitative stream allowed for 50 in-depth day long interviews with selected companies trialing the Assessment list on use-cases. Finally, the last channel allowed for feedback from the broader community, discussion papers, white papers, blog posts and reports were provided alike from entities as broad as individual researchers to international industry.

2020) is the first tool in the EU that took 'trustworthy AI' into account throughout an AI system's lifecycle and outlined how an assessment of that could take shape. It was also among the earliest serious attempts to translate ethical principles for AI into actionable measures for all stakeholders involved throughout the AI lifecycle, be that researchers, industry, government or civil society.

The concept of 'trustworthy AI', in the way that the AI HLEG defined it, became a cornerstone for AI governance in the EU. Building on its previous emphasis on 'human-centric AI' (as we have seen in the European Council's call to the European Commission,[52] the 'Declaration' (European Commission, 2018c) and the 'Coordinated Plan' (European Commission, 2018b)) the European Commission adopted this conceptual approach and expanded upon it in the Communication on 'Building Trust in Human-Centric AI' (European Commission, 2019). In that Communication, the European Commission supports the key requirements and the concept of trustworthy AI, stating that;

> "Only if AI is developed and used in a way that respects widely-shared ethical values, it can [sic] be considered trustworthy."

This can be understood – and as we will see, is reflected in subsequent governance documents and decisions – as the European Commission embracing the concept of 'trustworthy AI' as a core component of its strategic vision.[53]

Equally, it doubles down on the reputation of the European Union as a region that produces "safe and high-quality products" (European Commission, 2019). To consolidate its place as a leader on 'trustworthy AI' on an international stage, the Communication on 'Building Trust in Human-Centric AI' (European Commission,

[52] See: https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf.
[53] It should be noted that the key requirements and Ethics Guidelines are of a non-binding format.

2019) furthermore launched a consensus building an 'International Alliance for a human-centric approach to AI'.[54] Its goal is to share the EU's vision and ambitions with like-minded international partners.

Beyond that, the Communication on 'Building Trust in Human-Centric AI' (European Commission, 2019) further expands on elements of documents such as the 'Coordinated Plan' (European Commission, 2018b). It reiterates the core foci to boost the ecosystem, such as an increase in joint ventures, pooling of data and other building blocks for AI, as well as the strengthening of synergies across Member States. It proposed to launch a set of networks of AI research excellence centers under the Horizon 2020 research and innovation framework programme, to set up networks of AI-focused Digital Innovation Hubs (DHIs) and to develop and implement a model for data sharing and common data spaces amongst Member States and other stakeholders. These suggestions directly shape the current state of affairs as we will see in Section 2.

2. The Present

The third way: the EU's AI northstar

We have now reviewed the recent historical backdrop of the current EU's regulatory strategy, discussing both its roots and predecessors. This brings us to today. The EU is looking to develop an attractive alternative to US and Chinese approaches to AI governance. In order to gain a bird's-eye view of that third way, this section will highlight a select number of important and current developments, illustrating how they contribute to the EU's direction.

[54] See: https://digital-strategy.ec.europa.eu/en/funding/international-alliance-human-centric-approach-artificial-intelligence.

On 19th February 2020, the European Commission published a comprehensive package consisting of: the 'European Strategy for Data' (European Commission, 2020a), the report on 'Safety Liability and Implications of AI, the Internet of Things and Robotics' (European Commission, 2020f), and the 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (henceforth: White Paper on AI; European Commission, 2020d). Due to the limited scope of this chapter, we will focus on the 'White Paper on AI' (European Commission, 2020d).

This 'White Paper on AI' (European Commission, 2020d) followed European Commission president von der Leyen's promise in her political agenda to put forward "legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence."[55] It solidified the commitment to human-centric and 'trustworthy AI', adding the first step towards a future legislative framework built on the concept of 'trustworthy AI' to the new EU AI governance portfolio. The 'White Paper on AI' (European Commission, 2020d) is divided into two main sections, one on an Ecosystem of Trust, focusing on the first proposal for a regulatory framework, and one on an Ecosystem of Excellence, focusing on supporting the European AI ecosystem. Both of these closely match ambitions outlined in previous policy documents in Section 1, such those in Europe's AI strategy (European Commission, 2018a), and fit in with other recent governance efforts. I will therefore use the 'White Paper on AI's' (European Commission, 2020d) duality of policy and infrastructure to highlight how far the EU AI policy has come in each area since this chapter's introductory section and how they build on one another to make the EU a hub for 'trustworthy AI'.

On 21st April 2021, the European Commission published its package on a European approach for AI containing: a 'Communication on Fostering a European Approach to Artificial Intelligence' (European Commission, 2021a); a 'Coordinated Plan on AI: 2021 review' (European Commission, 2021b); and, the highly anticipated proposal

[55] See: https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf.

for a 'Regulation on a European approach for Artificial Intelligence (AI Act)' (European Commission, 2021c). The 'Coordinated Plan on AI: 2021 review' (European Commission, 2021b) builds on the 'Coordinated Plan' (European Commission, 2018b) and dramatically expands its scope and ambitions, and the 'Regulation on a European approach for Artificial Intelligence (AI Act)' (European Commission, 2021c) builds on the 'White Paper on AI' (European Commission, 2020d) and subsequent impact assessments conducted by the European Commission.[56] We will now see how all of this shapes up to form the context and the ecosystem for future AI governance in the EU.

2.1. Trust and the EU's AI governance

The chapter in the 'White Paper on AI' (European Commission, 2020d) dedicated to the Ecosystem of Trust outlined the European Commission's policy proposals for a potential regulation prior to the final publication of the proposal for a 'Regulation on a European approach for Artificial Intelligence (AI Act)' (European Commission, 2021c) on 21st April 2021. The chapter was strongly inspired by the conceptual idea of 'trustworthy AI' and heavily referenced the work of the AI HLEG.

The core proposal suggested that in an envisioned horizontal legislation mandatory legal requirements should apply to high-risk cases of AI only. These high-risk AI cases were defined by the following cumulative criteria: if the sector itself is high risk (e.g. healthcare, transport and if the intended use involves high risk (e.g. injury, death, significant material/immaterial damage).

The mandatory legal requirements largely reflect the 'Ethics Guidelines' (AI HLEG, 2019) seven key requirements and are composed of the following: a requirement for adequate training data; a requirement for data and record keeping; a requirement

[56] See: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=PI_COM:Ares(2020)3896535&from=EN.

for the provision of information; a requirement for robustness and accuracy; and, a requirement on human oversight. The final requirement was specifically laid out for the case of remote biometric identification. High-risk AI systems would be subject to conformity assessment (e.g. testing, inspection and certification) accounting for these requirements before they would be able to enter the EU market.

The 'White Paper on AI' (European Commission, 2020d) also outlined strategies for non-high-risk AI systems. It was suggested that these could partake in a voluntary labeling scheme which could build upon or implement the 'Assessment List' (AI HLEG, 2020). We see that the building blocks and vision sketched in Section 1 are starting to take considerable shape building the EU's AI governance future.

Soon after, the legal affairs committee of the European Parliament adopted several aligned reports. These tackled an ethical framework for AI, civil liability claims against operators of AI systems and the protection of intellectual property rights with regards to AI.[57] It is noteworthy that the first report's guiding principles strongly resembled those of the 'Ethics Guidelines'. We can infer that the vision of EU AI governance is coherent across the EU institutions, which is important[58] as we move to the most recent and high-profile policy development on the European Commission's side: the 'Regulation on a European approach for Artificial Intelligence (AI Act)' (European Commission, 2021c).

Following the publication of the 'White Paper on AI' (European Commission, 2020d), the European Commission conducted impact assessments and opened a stakeholder consultation to receive feedback on the 'White Paper on AI' (European

[57] See: https://www.europarl.europa.eu/news/en/press-room/20200925IPR87932/making-artificial-intelligence-ethical-safe-and-innovative. The second legislative initiative is on 'liability for AI causing damage', focusing on civil liability claims against AI-systems, and the third report addresses intellectual property rights (IPRs) with relation to AI, suggesting that AI lacks a legal personality, and therefore inventorship should be exclusive to humans.

[58] The proposal for a Regulation on a European approach for Artificial Intelligence will need to pass through the European Parliament and the Council. Once these two institutions agree on a final text, the regulation will be adopted.

Commission, 2020d). This feedback shaped the subsequent proposal for a regulation.

The proposed 'Regulation on a European approach for Artificial Intelligence (AI Act)' (henceforth: 'AI Act'; European Commission, 2021c) introduces the European Union's legislation for AI, specifically high-risk AI systems. It is a risk-based regulation which covers stand alone AI systems that are considered high-risk which are elaborated on in Annex III to the 'AI Act' and cover use cases such as in law enforcement for individual risk assessment, education and vocational training for determining access to educational or training institutions or specific cases of access to essential public and private services and benefits. In short, Annex III lists a number of areas and specific use cases in those areas where stand-alone AI systems will automatically be considered high risk due to their potentially adverse impact on health, safety or fundamental rights of persons or groups. The other case of high-risk AI systems are those that are not stand-alone AI systems but those that are safety components of products or systems, or those that are products or systems.

Both of these types of high-risk AI systems need to comply with a number of requirements the 'AI Act' (European Commission, 2021c) lays down in Title III Chapter II, although the manner in which that compliance is achieved, documented and assessed (conformity assessment) is different between stand-alone and integrated high-risk AI systems. In the case of high-risk AI systems that are safety components of products or systems, or are themselves products or systems, the harmonized 'AI Act' (European Commission, 2021c) adjusts to fit with the existing sectoral procedures, rules and regulations.

The scope of this 'AI Act' (European Commission, 2021c) encompasses a range of actors: providers that place their AI system on the EU market, users of AI systems in the EU (except those that use it in a personal, non-professional activity) and providers and users of AI systems that are not based in the EU but where the output of their AI system is used in the EU.

All high-risk AI systems need to fulfill the requirements set out in the 'AI Act' (European Commission, 2021c) Title III Chapter II. These requirements closely match those that were previously proposed in the 'White Paper on AI' (European Commission, 2020d) and, as we have seen, are therefore closely connected to the requirements within the 'Ethics Guidelines' (AI HLEG, 2019). The requirements listed in the 'AI Act' are: *Data and Data Governance*; *Technical Documentation*; *Record Keeping*; *Transparency and Provision of Information to Users*; *Human Oversigh*t; and *Accuracy, Robustness and Cybersecurity*.

Whilst they are not described in this format, I would like to propose that the requirements can be thought of in two categories: those that are procedural and those that are informative. Data and Data Governance, Human Oversight and Accuracy, Robustness and Cybersecurity are procedural. They concern themselves with the workings of the algorithm throughout its lifecycle and how these can be affected in a positive manner to avoid negative impacts. By contrast, Technical Documentation, Record Keeping and Transparency and Provision of Information to Users can be considered as informational requirements. They track procedural information, check it and monitor it throughout the AI system's life cycle.

Those actors that are responsible for ensuring that a high-risk AI system complies with the 'AI Act' have to fulfill certain conditions on top of adherence to the requirements mentioned previously. In short, they have to first build their AI system in accordance with the requirements from Title III Chapter II, then they have to undertake an internal conformity assessment of the AI system which encompasses paper trails and documentation generated in the first step and developed as a framework for the AI system throughout its functioning. That will entail a Quality Management System which ensures compliance with the 'AI Act' (European Commission, 2021c), a Risk Management system which acts as a continuous iterative process throughout the AI system's lifecycle, and Technical Documentation which covers elements such as detailed descriptions,

pre-determined changes of the AI system and the performance, as well as monitoring, functioning and control of the AI system. Third, the provider or relevant other actor needs to establish a post-market monitoring system for the AI system once it has been put on the market or placed into service. This post-market monitoring system will collect logs produced by the AI system, act as a supervisor to the AI system and report serious incidents if they occur. Finally, before the AI system can be put on the market or placed into service it must be registered in the EU database, and EU Declaration of Conformity must be filled out to describe its adherence to the 'AI Act' (European Commission, 2021c) and it should be affixed with a CE marking to indicate that it has passed its conformity assessment.

In addition to requirements and procedures for high-risk AI systems, the 'AI Act' (European Commission, 2021c) also lays down a number of AI systems that are prohibited for use in the EU under certain conditions. Without enumerating them in detail, these prohibited AI system cover those that deploy subliminal techniques that could cause harm, those that exploit vulnerabilities in a manner that would cause harm and those that public authorities could use to evaluate the trustworthiness of an individual, leading to unfavorable treatments in different contexts or treatment that is disproportionate. Moreover, it includes 'real-time' biometric identification systems if they are used in publicly accessible spaces and for the purpose of law enforcement. However, noteworthy exceptions to the latter are cases where there is a targeted search for potential victims of crime, where it is in the public interest to prevent specific, substantial and imminent threats and to detect certain perpetrators.

Akin to the proposals in the 'White Paper on AI' (European Commission, 2020d), the 'AI Act' (European Commission, 2021c) also briefly concerns itself with voluntary Codes of Conduct for non-high risk AI systems, with the intention to foster 'trustworthy AI' and, therefore, compliance to the 'AI Act' within the broader ecosystem. The next subsection will discuss how the corresponding environment within the EU is boosted in order to establish the overarching framework that these

policy and legislative ambitions would fit in with. Section 3 will then concern itself in more detail with specific elements of the ecosystem that are likely to become crucial to the EU's success in AI governance in the near future.

2.2. Strengthening the AI ecosystem

In recent years, it has become clear that the EU does not solely wish to rely on their regulatory expansionism, exporting norms and legislative approaches towards 'trustworthy AI' on an international stage. Acknowledging that its ecosystem has, at times, difficulty competing with tech giants developing or established outside of the EU, it is increasingly moving towards Digital Sovereignty. This encompasses the broader AI landscape in the EU. In order to have a truly comprehensive and integrated approach towards AI governance, ethical, policy and regulatory efforts must be boosted in tandem with the existing and foreseen landscape. In short, an increase in relevant EU infrastructure for AI development, deployment and use, equals an increase in ownership of the technology, an increase in the ability to shape it directly through norms for trustworthy AI (through soft and hard law) and a decrease in reliance on outside actors. Keeping this in mind, the following paragraphs will sketch how the EU is building this infrastructure and what benefit this may yield, starting with the chapter in the 'White Paper on AI' on an Ecosystem of Excellence and expanding it further with efforts outlined in the 'Coordinated Plan on AI: 2021 review' (henceforth: 'Coordinated Plan: 2021 review'; European Commission, 2021b) and adjacent initiatives.

Many of the areas below directly link back to aspects mentioned in Section 1 such in the 'Declaration' (European Commission, 2018c), the 'Coordinated Plan' (European Commission, 2018b) and in the AI Strategy (European Commission, 2018a), which all outlined the need to boost the ecosystem, to combine resources and to increase

technical capabilities to ensure the EU's leadership in human-centric and 'trustworthy AI'.

The 'White Paper's' (European Commission, 2020d) chapter on an Ecosystem of Excellence concerns itself with technical infrastructure as well as with ecosystem building. On the latter, it especially focuses on building new infrastructures. On the more research-oriented side, it proposed work on establishing a lighthouse center of research, innovation and expertise, to combat a seemingly fragmented AI research community in the EU. Looking at industry, small and medium-sized enterprises (SME) and start-ups, it calls for the development of testing and experimentation facilities (TEFs), building out capacity via the Digital Innovation Hubs (as mentioned in the 'Coordinated Plan'; European Commission, 2018b), a recently funded AI-on-Demand platform[59] and engagement of key stakeholders through a Public-Private Partnership on AI, data and robotics in the context of Horizon Europe.[60] It also touches on turning the 'Assessment List' (AI HLEG, 2020) into an indicative curriculum for those developing AI, and an ambition to keep talent and attract talent to the EU through new education networks under the Digital Europe programme.[61]

In accordance with the 'European Data Strategy' (European Commission, 2020a), the 'White Paper on AI' also puts an emphasis on the role of data in AI development ("compliance of data with the FAIR principles will contribute to build trust and ensure re-usability of data"; European Commission, 2020d), as well as the importance of computing infrastructure. This will be explored in more detail at the end of this section.

[59] See: https://cordis.europa.eu/project/id/825619.
[60] Horizon Europe is the current Multiannual Financial Frameworks programme and runs from 2021-2027 to support research, science and innovation with EUR 95.5 bn.
[61] See: https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

Every single one of these proposed efforts ties in with the 'Coordinated Plan: 2021 review' (European Commission, 2021b) and demonstrates the EU's efforts to build an infrastructure that can match its ambition on the governance side to promote the development of human-centric, sustainable, inclusive and trustworthy AI. While the 'Coordinated Plan' (European Commission, 2018b) mapped out the initial areas in which the Member States should pool their resources and coordinate their actions, the 'Coordinated Plan: 2021 review' (European Commission, 2021b) moves towards an action-oriented approach with concrete joint actions focusing on the implementation of concrete measures and the removal of remaining fragmentation.

It is built around four key pillars: (1) to set enabling conditions for AI development and uptake; (2) to make the EU a place where excellence thrives from the lab to the market; (3) to ensure that AI works for people and society as a force for good; and (4) to build strategic leadership in high-impact sectors. These high-impact sectors encompass areas such as smart mobility, law enforcement, migration and asylum, climate and the environment. In tandem with the third pillar, which encompasses a promotion of 'trustworthy AI' globally and nurturing of talent and skills, it could be seen as a reflection of the second component in the 'Communication on AI in Europe, that is the EU's initial AI strategy, which focused on socioeconomic changes associated with AI. Although the focus on scaling up the EU technical infrastructure is evidenced across all four pillars, the first is of particular relevance.

The policy document homes in on governance coordination frameworks and, crucially, on data infrastructures and computing capacities in order to create an enabling environment for AI development and uptake.

Referring back to the 'European Strategy for Data' (European Commission, 2020a), which aims to establish a single market for data within the EU and and the 'Proposal for a regulation on European data governance (Data Governance Act)' (European Commission, 2020e), which proposes several regulatory measures to

increase society's trust in data sharing, the 'Coordinated Plan: 2021 review' (European Commission, 2021b) outlines a number of core actions for data and an associated cloud infrastructure. These actions include establishing a new European Alliance for Industrial Data, Edge and Cloud,[62] co-investing with the Member States in common European data spaces and a European cloud federation and investigating the opportunity to set up an Important Project of Common European Interest (IPCEI) for next generation cloud infrastructures.

Prior to the 'Coordinated Plan: 2021 review' (European Commission, 2021b), at the end of 2020, 27 Member States signed a joint 'Declaration on Building the next generation cloud for businesses and the public sector' (European Commission, 2020c) where they expressed their intention to establish a secure, trustworthy and competitive cloud infrastructure in Europe for public administration, businesses and citizens alike. It addresses efforts aligned with those in the 'Coordinated Plan: 2021 review' (European Commission, 2021b), such as pooling of EU, national and private investment and shaping the process in accordance with the European Alliance on Industrial Data and Cloud,[63] and fostering technical solutions and policy norms for an interoperable pan-European cloud service.

Adjacent to this is Gaia-X,[64] an independent European platform for cloud infrastructure launched by France and Germany[65] intends to increase European cloud competitiveness vis-a-vis the US and China.

With an eye to infrastructure, the 'Coordinated Plan: 2021 review' (European Commission, 2021b) looks to support the development of High Performance

[62] See: https://digital-strategy.ec.europa.eu/en/library/cloud-and-edge-computing-different-way-using-it-brochure.
[63] See: https://digital-strategy.ec.europa.eu/en/news/towards-next-generation-cloud-europe.
[64] See: https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html.
[65] See: https://www.euractiv.com/section/digital/news/digital-brief-the-gaia-x-generation/?utm_content=1591278775&utm_medium=eaDigitalEU&utm_source=twitter.

Computing capabilities, as well as AI hardware. The latter encompasses investment in micro-electronics for AI chips, neuromorphic computing, photonics and projects under the Electronic Components and Systems for European leadership Joint Undertaking (ECSEL JU).[66] More specifically, actions call for the launch of an Industrial Alliance on Microelectronics, supporting research and innovation actions for low-power edge AI, and investing in processor and semiconductor technologies.

In fact, as part of its goal of achieving Digital Sovereignty, the EU quite evidently is aiming to advance its capabilities and to lessen its reliance on international actors when it comes to the design and production capabilities of low-power processors for AI and towards 2nm processor technologies. In late 2020, 18 Member States signed a 'Declaration on a European Initiative on Processors and Semiconductor Technologies' (European Commission, 2020b) to consolidate resources and boost the EU's electronics and embedded systems value chain.

The 'Coordinated Plan: 2021 review' (European Commission, 2021b) also accounts for High-Performance Computing. It encourages Member States to continue developing large-scale High-Performance Computing infrastructure and references the importance of the EuroHPC Joint Undertaking.[67]

A proposed new independent Regulation for the EuroHPC[68] is expected to lead to an increase in the acquisition and development of supercomputers in the EU, rebalancing the scale in favor of the EU. Overall, it aims to develop exascale supercomputers with over '1 billion billion' operations per second (10^18 ops/second), to support the development of quantum and hybrid computers (also described in the 2018 regulation, section 21) and to create 33 'national competence

[66] See: https://www.ecsel.eu/what-we-do-and-how#:~:text=The%20ECSEL%20Joint%20Undertaking%20%2D%20the,era%20of%20the%20digital%20economy.
[67] See: https://eurohpc-ju.europa.eu/.
[68] See: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1592.

centers' which will help to provide easier access to HPC opportunities locally and strengthen knowledge and expertise.

The second pillar in the 'Coordinated Plan: 2021 review's' (European Commission, 2021b) focuses on knowledge transfer and horizontal actions to support research and innovation (R&I). Its actions cover stakeholder collaboration, expanding and mobilizing research capacities, building up suitable TEFs and funding AI solutions and ideas. Stakeholder collaboration will range from the Public-Private Partnership on AI, Data and Robotics[69] to a co-programmed European Partnership on Photonics,[70] supporting the EU's drive towards technological sovereignty. Whereas the European Commission already invested over €50 million in AI excellence centers through Horizon 2020,[71] it suggests funding more AI excellence centers and encourages Member States individually to set up regional and national excellence centers. Reminding ourselves of this chapter's earlier section on the importance of 'trustworthy AI' for EU AI governance, it needs to be highlighted that the document suggests that funded programmes for AI under Horizon Europe are expected to adhere to the 'ethics by design' principle, including 'trustworthy AI'. We can see that various earlier threads are starting to come together. Finally, as mentioned in the 'White Paper on AI' (European Commission, 2020d) and the earlier 'Coordinated Plan' (European Commission, 2018b) Digital Innovation Hubs (DIHs) play a role in strengthening the ecosystem. To that end, alongside TEFs for specific sectors, such as edge AI or agri-food, the European Commission will support a scaling up of existing DIHs and set up new networks for what it terms European Digital Innovation Hubs (EDIHs) with AI expertise. These EDIHs will connect SMEs and start-ups with resources made available via the AI-on-Demand platform and relevant TEFs to make the AI system ready for deployment within the EU market.

[69] See: https://ai-data-robotics-partnership.eu/.
[70] See: https://www.photonics21.org/#:~:text=The%20European%20Technology%20Platform%20Photonics21, growth%20and%20jobs%20in%20Europe.
[71] See: https://digital-strategy.ec.europa.eu/en/news/towards-vibrant-european-network-ai-excellence.

Finally, an investment of €1 billion from Horizon Europe and the Digital Europe programmes is expected between 2021-2027. The Digital Europe programme overall budget funds artificial intelligence (€2.1bn), HPC (€2.2bn) and cybersecurity (€1.7bn). The ambition remains the same as in the earlier 'Coordinated Plan' (European Commission, 2018b), namely to raise this to €20bn per year through public and private investment. Other institutions that are expected to fund AI in the EU are the European Innovation Council,[72] the European Investment Bank[73] (via the European Innovation Fund[74]) and the European Institute of Innovation and Technology.[75]

One issue the EU has historically faced is that promising companies are often purchased by foreign companies before they reach their full potential, gobbling up talent, and knowledge in the process. Although this is not an explicit part of strengthening the EU's AI landscape nor mentioned in the 'Coordinated Plan: 2021 review' (European Commission, 2021b) it is relevant to quickly mention the EU's regulatory framework to screen foreign direct investment (FDI).[76] This FDI framework aims to protect the EU's strategic interests and came into force at the end of 2020. Of particular relevance, in light of the EU's shift towards achieving digital sovereignty[77] and scaling its technical infrastructure, is that this framework covers assets that are 'critical technologies and dual use items'.[78] This includes amongst others AI, robotics and semiconductors.

From Sections 2.1 and 2.2 it is evident that the EU is both (1) serious in its pursuit to strengthen its vision of human-centric 'trustworthy AI' by shaping the AI governance framework through regulation, policy and certification; and, (2) willing

[72] See: https://eic.ec.europa.eu/index_en.
[73] See: https://www.eib.org/en/index.htm.
[74] See: https://ec.europa.eu/clima/policies/innovation-fund_en.
[75] See: https://eit.europa.eu/.
[76] See: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019R0452.
[77] See: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en.
[78] As defined in Article 2.1 of Regulation (EC) No 428/2009.

to build out the entire ecosystem in support of this vision, positioning itself as a future sovereign digital actor and third way between the US and China.

3. The Future
Sketching the future of AI governance in the EU

In the previous two sections we saw how the EU built up its strategy and how all of the elements fit in with the larger tapestry of the EU's approach to AI governance. Section 3 will be based on the current dynamic that the EU is exhibiting and briefly sketch three AI governance areas that are prime candidates to become crucial for AI governance in the EU in the coming decade.

The finalization and implementation of the 'AI Act' (European Commission, 2021b) over the coming months and years is a clear candidate, but there are other less obvious but equally relevant ones. The following paragraphs pick them out, polish them and highlight their importance in the future pathways for EU AI governance.[79]

3.1. AI Megaprojects: a CERN for AI and AI lighthouses

Multiple experts have called for megaprojects within the EU over the past few years. Most notably, for a CERN for AI.[80] Certainly, such a project would be very ambitious. Nevertheless, upon careful reading of recent EU policy documents and the general drive to boost the technical landscape and ecosystem (as outlined in Section 2.2) a budding AI megaproject may be on the cards.

[79] It should be noted that this section is the personal opinion of the author and the likelihood of the proposed sketches coming into fruition varies.
[80] See: https://www.steven-hill.com/why-we-need-a-cern-for-ai/.

There are roughly two shapes such a project could take: being centralized within a Member State who has suitable technical infrastructure and is well located or broadly distributed across Member States, with one centralized headquarter. Given the existing ecosystem, ongoing efforts to boost research capacity and technical infrastructure and recent advocacy from large research groups within the EU,[81] both options may be viable.

As described earlier, the EU plans to build out their existing DIHs into EDIHs with significant focus on AI EDIHs. This would lead to a stark increase in the number of facilities where research can be conducted and where AI systems can be developed and experimented upon by SMEs and start-ups. Moreover, the European Commission has recently funded a large scale project called ELISE,[82] the European Network of AI Excellence Centers. ELISE collaborates with the European Laboratory for Learning and Intelligent Systems (ELLIS), a large network of European researchers, and closes the gaps between AI institutes in Europe. This adds to a previously funded network, TAILOR[83] (Foundations of Trustworthy AI - Integrating Learning, Optimization and Reasoning) whose goal is to create a network across Europe on the "Foundations of Trustworthy AI".

Efforts such as these evidence that there is fertile ground to establish a centralized large-scale headquarter from an increasingly powerful network and quasi-independent nodes.

Another key reason for why a large-scale AI project might be both on the cards and meaningful for the EU to establish can be found in the 'AI Act' (European Commission, 2021b). In tandem with new regulatory requirements, more TEFs will

[81] See: https://claire-ai.org/wp-content/uploads/2020/02/CLAIRE-Press-Release-11.pdf; https://www.timeshighereducation.com/news/scientists-split-europe-paves-way-cern-of-ai; https://sciencebusiness.net/news/call-cern-ai-parliament-hears-warnings-risk-killing-sector-over-regulation.
[82] See: https://cordis.europa.eu/project/id/951847.
[83] See: https://liu.se/en/research/tailor/about.

be needed. This ranges from TEFs specialized to test and assess for specific aspects of the conformity assessment, as well as those that can assess an AI system's entire regulatory fitness. Building out existing facilities for testing and experimentation and establishing novel ones will eventually lead to a dense landscape of distinct but similar institutions across the EU's Member States. It might be in the EU's best interest to centralize these facilities and locate them alongside big industrial efforts such as the European Cloud efforts, European Data Spaces and the HPC Joint Undertaking. Such a localization could: increase efficiency and provide economies of scale for using data, research engineering, and other supporting infrastructure; enable more ambitious research, testing and experimentation efforts; and, encourage a lasersharp alignment between policy and practice.

Indeed, the European Commission has indicated ambitions to develop something akin to a CERN for AI in several policy documents, for example in the 'White Paper on AI' (European Commission, 2020d), and most recently in the 'Coordinated Plan on AI: 2021 review' (European Commission, 2021b). In particular, the development of AI Lighthouse Centers (or, a center) within the EU is championed. This would be a large-scale research facility for AI.

It would be promising for the EU's ambitions on a global playing field to establish an AI lighthouse center, a CERN for AI or another version of a large-scale facility for AI research and development. Most importantly, this could lead to the EU becoming a truly unified player where fragmentation between various European research institutes is superseded (Stix, 2018), significant chunks of the aimed for EUR 20bn per year funding for AI could be centralized for ambitious projects, and new talent could be attracted to the EU (Stix, 2019).

Considering a future landscape with an increasing number of networked institutions, mounting calls from large AI research networks within the EU, and the

policy proposals from the European Commission, the future of EU AI governance may well hold an AI megaproject.

3.2. AI Agencies: regulation, measurement and foresight

The idea of a large European AI Agency is not new. The very first document presented in this chapter, the resolution on 'Civil Law Rules on Robotics' (European Parliament, 2017), already called for the establishment of an EU Agency for Robotics and AI in "order to provide the technical, ethical and regulatory expertise needed to support the relevant public actors, at both Union and Member State level, in their efforts to ensure a timely, ethical and well-informed response to the new opportunities and challenges, in particular those of a cross-border nature".

Similarly, the 2019 European Parliament report 'A comprehensive European industrial policy on artificial intelligence and robotics' (European Parliament, 2019) called for the establishment of a European regulatory agency for AI and algorithmic decision-making. With this backdrop, and tracing the institutional landscape mapped out by the 'AI Act' (European Commission, 2021c) it is likely that the EU will eventually establish a new institution, specifically for the governance of AI. The 'AI Act' (European Commission, 2021c) envisions a complex institutional interplay to sustain the regulatory measures for AI. This encompasses various national institutions: those that fall under the National Competent Authorities, which would be the National Supervisory Authority, the Notifying Authority and various Notified Bodies (official conformity assessment bodies); Market Surveillance Authorities; and, from the European Commission's side a novel European AI Board (where e.g. member states will be represented) and an expert group. All of these will play a crucial role for the application of the horizontal regulation for AI within the EU and will have different scopes and powers. Some will have investigative power and some will assess the suitability of AI systems for the European market. Of course, many of these institutions cannot (and should not) be merged. Nevertheless, after an

initial phase of getting to know the ropes of the final agreed upon regulation, it is likely that there will be a time window in which an EU AI Agency would be built to combat fragmentation, pool expertise and streamline various workflows.

Beyond the aforementioned scopes, the 'AI Act' (European Commission, 2021c) also has a provision which ensures that new AI systems can be added to the list of high-risk AI systems as and when deemed appropriate. In order to ensure that timeliness and foresight are underlining this power, and more generally, to ensure that policy making matches technological progress, I propose that another version of an EU AI Agency -- which has not been part of any EU-level discussions yet -- should be considered: a European AI observatory.

Historically, observatories were established to measure and survey natural occurrences e.g. astronomical, geophysical or meteorological events. An AI observatory as envisaged here, on the other hand, would monitor, measure and benchmark AI progress, a technology created by humans.

Although the EU is involved in OECD efforts towards an international AI policy observatory and has its own body, the AI Watch,[84] this does not yet live up to what an EU AI observatory could look like. As envisaged here, a EU AI observatory should have the capacity and ambition of conducting independent forecasting and measurement exercises. These in turn would ensure that policy making, regulation and other governance efforts in the EU are aligned with the technical state of the art of AI systems and sufficiently future proof.

As previously indicated, in order for policymakers and regulators to make suitable and timely decisions to add potential future high-risk AI systems to the 'AI Act' (European Commission, 2021c), either to regulate them or to ban them, they need to

[84] A joint initiative between the European Commission's Joint Research Center (JRC) and the Directorate General for Communications Networks, Content and Technology (DG CONNECT). See more: https://knowledge4policy.ec.europa.eu/ai-watch/about_en.

be aware of ongoing technical developments. One way of doing so from a government's perspective could be to monitor the technical landscape, measure technical progress, and therewith notice crucial shifts that could indicate a cause for concern or intervention.

Overall, AI governance in the EU, through regulation, standards, certification or other efforts could be significantly more impactful, agile and anticipatory if it narrowed the pacing gap between technological progress and governance efforts (Marchant et al., 2011). Furthermore, metrics can be seen as comparatively non-threatening and could encourage information sharing between countries and institutions, indirectly promoting collaboration and cooperation. Taking these aspects into consideration, I suggest that an EU AI observatory would be vastly beneficial to the EU's AI ambitions and could be seen as a contributing factor for better regulatory measures in the future.

3.3. Standards

Lastly, standards will play an important role in the future of the EU's AI governance. While standardization efforts are not a 'European-only' effort, they are likely to meaningfully shape the EU market for AI systems. Crucially, the 'AI Act' (European Commission, 2021c) notes that if suitable standards exist[85] that would cover one, or more of the relevant legal requirements for an AI system to pass conformity assessment (outlined in Section 2.2.) then an adherence to those standards can be considered as an adherence to the legal requirement(s) in question. Of course, if a provider chooses not to follow an existing standard or, where a standard does not exist then they must prove suitable and sufficient adherence to the legal obligations in a different manner. This goes to say that actors involved in standardization bodies and those directly working on standards will

[85] Those standards would have to be published in the Official Journal of the European Union.

have some non-negligible leeway in shaping the future mechanisms with which many of those adhering to the regulatory framework for AI will tackle their conformity assessments. Standardization efforts might framing some of the hurdles a high-risk AI system needs to pass before it enters the EU market and shape the manner in which relevant actors think about assessing high-risk AI systems.[86] I propose that it is a key lever for upcoming governance measures within the EU.

4. Conclusion

In conclusion, this chapter first introduced the background of the EU's AI governance ambitions, drawing together the different elements and highlighting how they interconnect, together developing the tapestry out of which the EU's vision and current governance efforts result. Subsequently, it introduced the two-pronged recipe the EU pursues for AI. There, it discussed the corresponding roles of 'trustworthy AI' and regulatory efforts with the associated scaling of the technical infrastructure within the EU. Together, it demonstrated how these choices support the EU's ambition to be a leader on ethical and human-centric AI on an international stage. Finally, the author sketched three possible future directions they envision the EU moving towards: AI research megaprojects, new AI Agencies, and an increasing importance of standardization efforts.

[86] Assuming the provider chooses to use standards or technical specifications for their conformity assessment. However, if standards are available it is likely that most providers will choose to adhere to the standards to streamline and minimize their workflows between various geographical regions.

Bibliography

European Commission. (2018a). *Communication from the Commission - Artificial Intelligence for Europe* (COM(2018) 237 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

European Commission. (2018b). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Coordinated Plan on Artificial Intelligence* (COM/2018/795 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN.

European Commission. (2018c). *(Digital Day) Declaration on Cooperation on Artificial Intelligence*, European Commission website - JRC Science Hub - Communities, https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/eu-declaration-cooperation-artificial-intelligence.

European Commission. (2019). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human-Centric Artificial Intelligence* (COM/2019/168 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168.

European Commission (2020a) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - A European strategy for data* (COM(2020) 66 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC00

66&from=EN.

European Commission. (2020b). *Declaration A European Initiative on Processors and semiconductor technologies*, European Commission website - Library, https://digital-strategy.ec.europa.eu/en/library/joint-declaration-processors-and-semiconductor-technologies.

European Commission. (2020c). *Declaration - Building the next generation cloud for businesses and the public sector in the EU*, European Commission website - News & Views, https://digital-strategy.ec.europa.eu/en/news/towards-next-generation-cloud-europe.

European Commission. (2020d). *On Artificial Intelligence - A European approach to excellence and trust* (COM(2020) 65 final), Brussels: European Commission, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Commission. (2020e). *Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act)* (COM/2020/767 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767.

European Commission. (2020f). *Report from the Commission to the European Parliament, the Council, and the European Economic and Social Committee - Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics* (COM/2020/64 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX:52020DC0064.

European Commission. (2021a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social*

Committee and the Committee of the Regions - Fostering a European approach to Artificial Intelligence (COM/2021/205 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:205:FIN.

European Commission. (2021b). *Coordinated Plan on Artificial Intelligence 2021 Review*, Brussels: European Commission, https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

European Commission. (2021c). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (COM/2021/206 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

European Economic and Social Committee. (2017). *Opinion of the European Economic and Social Committee on 'Artificial intelligence — The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society' (own-initiative opinion)* (2017/C 288/01), Brussels: Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016IE5369.

European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and 'autonomous' systems*, Brussels: Publications Office of the European Union, https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382.

European Parliament. (2017). *European Parliament resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics* (2018) *(*2015/2103(INL))(2018/C 252/25), Brussels: Official Journal of the

European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051&qid=1620812299497.

European Parliament (2019) *Report on a comprehensive European industrial policy on artificial intelligence and robotics* (2018/2088(INI)), European Parliament website, https://www.europarl.europa.eu/doceo/document/A-8-2019-0019_EN.html.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. https://doi.org/10.2139/ssrn.3518482.

Hagendorff, T. (2019). The Ethics of AI Ethics -- An Evaluation of Guidelines. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1903.03425.

Independent High-Level Expert Group on Artificial Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, Brussels: European Commission, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Springer Science & Business Media.

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

Stix, C. (2018). The European AI Landscape. *Workshop Report. Brussels: European Commission. DG Connect. Retrieved from Http://ec. Europa. Eu/newsroom/dae/document. Cfm*.

Stix, C. (2021). Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics*, *27*(1), 15.

Winfield, A. (2019). An updated round up of ethical principles of robotics and AI. *Retrieved on August 13th*.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1812.04814

Chapter III

Actionable principles for artificial intelligence policy: Three pathways

……..

Stix, C. Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics* 27, 15 (2021). https://doi.org/10.1007/s11948-020-00277-3.

……..

Introduction

Recent years have seen a veritable surge of Ethics Principles[87] for artificial intelligence[88] (AI) (Fjeld et al., 2020; Hagendorff, 2020; Ryan & Stahl, 2020; Jobin et al., 2019; Zeng et al., 2018; Morley et al., 2019). This in turn led to critical debates over the usefulness and impact of such instruments, with a particular focus on what is often held as a lack of implementation of such AI Ethics Principles into actual policy-making. In response, this paper will propose a preliminary framework to support and improve the implementation of AI Ethics Principles in governmental policy at this critical time.

The appeal of AI Ethics Principles lies in their promise to condense complex ethical considerations or requirements into formats accessible to a significant portion of society, including both the developers and users of AI technology. To live up to this, however, these principles face two high-level challenges: (a) they must achieve a

[87] For the purpose of this paper, "AI Ethics Principles" encompass all documents outlining policy for the development and deployment of AI, based on ethical considerations.

[88] For the purpose of this paper, the author follows the definition of AI of the European Commission (European Commission, 2018, p.1) stating that AI systems "display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals" .

succinct condensation of broad and deep ethical theories into an accessible number of principles, and (b) they must strike a balance between pursuing an ideal hypothetical outcome, and working to secure workable pragmatic outcomes. In doing so, it is important to recognize that while 'workable pragmatic outcomes' may rightly be perceived as suboptimal from a strict ethical perspective, they will often form the critical basis to moving AI Ethics Principles forwards into practical policy.

This paper focuses on helping AI Ethics Principles strike such a balance, by complementing previous work (which focuses largely on identifying the ideals to be pursued) with a perspective that enables the achievement of 'workable pragmatic outcomes' in AI policy. To do so, it proposes one avenue to increase AI Ethics Principles' operationalizability into policy, ensuring these are actionable for governmental actors. It therefore limits itself to exploring one particular bottleneck that AI Ethics Principles face, and does not claim to resolve other, equally pressing shortcomings. In short, the goal is not to explore what precisely should be *in* AI Ethics Principles per se, but examine one angle as to *how* they should be developed to be actionable in a specific domain.

This paper puts forward an initial framework for the development of what it calls *Actionable Principles for AI Policy*. To do so, it will proceed as follows: in section "Case Study: The Ethics Guidelines for Trustworthy Artificial Intelligence" it will present three key procedural elements of the 'Ethics Guidelines for Trustworthy Artificial Intelligence' (henceforth: 'Ethics Guidelines', AI HLEG, 2019b) presented in 2019 by the European Commission's independent High Level Expert Group on Artificial Intelligence (AI HLEG). Reviewing these procedural instruments, this paper will subsequently expand and build thereon. On this basis, section "A Preliminary Framework for Actionable Principles" will culminate in a proposal for an initial framework for Actionable Principles for AI.

Actionable Principles for AI

In many areas, including AI, it has proven challenging to bridge ethics and governmental policy-making (Müller 2020, 1.3). To be clear, many AI Ethics Principles, such as those developed by industry actors or researchers for self-governance purposes, are not aimed at directly informing governmental policy-making, and therefore the challenge of bridging this gulf may not apply. Nonetheless, a significant subset of AI Ethics Principles are addressed to governmental actors, from the 2019 OECD Principles on AI (OECD, 2019) to the US Defense Innovation Board's AI Principles adopted by the Department of Defense (DIB, 2019). Without focusing on any single effort in particular, the aggregate success of many AI Ethics Principles remains limited (Rességuier and Rodriques 2020). Clear shifts in governmental policy which can be directly traced back to preceding and corresponding sets of AI Ethics Principles, remain few and far between. This could mean, for example, concrete textual references reflecting a specific section of the AI Ethics Principle, or the establishment of (both enabling or preventative) policy actions building on relevant recommendations. A charitable interpretation could be that as governmental policy-making takes time, and given that the vast majority of AI Ethics Principles were published within the last two years, it may simply be premature to gauge (or dismiss) their impact. However, another interpretation could be that the current versions of AI Ethics Principles have fallen short of their promise, and reached their limitation for impact in governmental policy-making (henceforth: policy).

It is worth noting that successful actionability in policy goes well beyond AI Ethics Principles acting as a reference point. Actionable Principles could shape policy by influencing funding decisions, taxation, public education measures or social security programs. Concretely, this could mean increased funding into societally relevant areas, education programs to raise public awareness and increase vigilance, or to rethink retirement structures with regard to increased automation. To be sure,

actionability in policy does not preclude impact in other adjacent domains, such as influencing codes of conduct for practitioners, clarifying what demands workers and unions should pose, or shaping consumer behavior. Moreover, during political shifts or in response to a crisis, Actionable Principles may often prove to be the only (even if suboptimal) available governance tool to quickly inform precautionary and remedial (legal and) policy measures.

There exist concrete examples demonstrating that some select AI Ethics Principles already do possess a degree of actionability. For instance, the Ethics Guidelines (AI HLEG, 2019b) have had a significant policy impact within the European Union (EU). They influenced both the political Agenda of Commission President Von der Leyen (2019) and informed the initial legislative framework proposal for AI. The latter used the Ethics Guidelines' seven key requirements for 'trustworthy AI'[89] to form the basis for the legal obligations which any 'high-risk' AI system would need to fulfill in order for it to be deployed within the EU (European Commission 2020). Arguably, these Ethics Guidelines were among the chief documents available to inform and guide the content of this legislative proposal. The aforementioned highlights the potential importance — and promise — of how future Actionable Principles could be set to significantly shape the development, deployment and use of AI by virtue of their influence on policy.

This paper will inevitably touch on a variety of concerns that afflict AI Ethics Principles beyond their lack of actionability. These are often intertwined with (if not the result of) procedural shortcomings that affect actionability. Some of these are, for example, a *lack of clarity* which can contribute to divergent interpretations (Whittlestone et al., 2019) or a *lack of balanced participation* which can contribute to 'ethics washing' (Floridi et al., 2018). The latter practice is commonly alleged with regards to industry-driven AI Ethics Principles, and can result in superficial

[89] Defined by the AI HLEG as being (i) ethical, (ii) lawful, and (iii) robust from a socio-technical perspective.

proposals that mask themselves as ethical, but may, in fact, be commercially or politically motivated (Floridi et al., 2018, p. 187).

Finally, it should be noted that underlying all this, there are a multitude of parallel discussions about the best governance approaches towards AI — and whether novel AI Ethics Principles are even the best tool in the first place. For instance, one prominent proposal is to use the international human rights framework as the basis for setting the ground for ethical AI systems. Indeed, this tool could be a valuable angle given the existing legitimacy and consensus this approach can draw from. Nevertheless, this paper takes a different angle and focuses on improving the ability of AI Ethics Principles to shape a given governance framework. It does not make a judgment on the relative value of various approaches in AI governance, but instead focuses on the improvement of one specific approach. Current AI Ethics Principles have been critiqued for the fact that, while they "may guide the entities that commit to them, […] they do not establish a broad governance framework." (Donahoe & Metzger 2019, p. 118). To that end, this paper proposes a concept of 'Actionable Principles'. Inevitably, some suggestions for Actionable Principles will draw on what has been successful in the past, and therefore stands to be promising in the future. These procedural elements, such as proposals for multi-stakeholder and cross-sectoral dialogue (Donahoe & Metzger, 2019; Yeung et al., 2019), are likely to overlap with, for example, ongoing suggestions for the establishment of an international human rights based framework for the purpose of AI governance. Building a framework for policy-effective Actionable Principles for AI is therefore not meant to be in competition with other approaches, but rather to complement them.

1. Case study: The Ethics Guidelines for Trustworthy Artificial Intelligence

This section focuses on the development process of the AI HLEG's Ethics Guidelines (2018–2019), in order to highlight three promising procedural elements, which will be subsequently developed in section "A Preliminary Framework for Actionable Principles." The Ethics Guidelines are selected because they arguably advanced the state-of-the-art of AI Ethics Principles by virtue of directly informing policy-making within the EU (European Commission 2020), and because they are grounded in the protection of fundamental rights (AI HLEG, 2019a, b; AI HLEG, 2020). The Ethics Guidelines therefore constitute a promising case to draw transferable or generalizable lessons from for Actionable Principles.

Of course, it has to be noted that neither the Ethics Guidelines nor their development process are void of criticism. Various critiques have been raised to it, from their alleged development under outsized industry influence (Hidvegi & Leufer, 2019; Article 19, 2019) and the obfuscation of 'red lines' (Metzinger, 2019; Klöver & Fanta, 2019), to a lack of matching governance structures to achieve real impact (BEUC, 2019; Veale, 2020). As such, it is important to note that although the framework for Actionable Principles proposed in this paper is informed by certain procedural elements of the Ethics Guidelines, it makes no assumption about the relative value of existing criticism of the Ethics Guidelines as a whole.

1.1. Diversity

The AI HLEG was a large group of experts from multiple sectors, ranging from ethicists, lawyers, to machine learning researchers, trade unionists, and various other stakeholders.[90] This diversity allowed the AI HLEG to provide informed recommendations sensitive to a variety of concerns. Moreover, the AI HLEG benefited from continuous engagement (including on earlier drafts of the Ethics Guidelines) with the European AI Alliance, a multi-stakeholder platform with over

[90] See: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

4000 entities and various subject experts across Europe.[91] Notwithstanding that, there is room for improvement in the establishment of groups developing Actionable Principles, especially in light of concerns over dominant representation of industry friendly voices (Hidvegi, Leufer, 2019) and a lack of sufficient representation of AI Ethicists (Metzinger, 2019) within the AI HLEG. While the European Commission has a strict selection process with different types of membership criteria during open calls,[92] it could, in the future, improve a lack of representation especially from civil society voices such as the European Network Against Racism, by explicitly hand-selecting them outside of public calls. Such a procedure could be within the remit of considering it an "overriding priority" that certain groups are adequately represented in given expert groups through membership.

1.2. Working methods

The AI HLEG solicited public feedback not once, but twice, both during the development process, but also after publishing the final document. The first public consultation concerned the draft Ethics Guidelines, made available for public feedback roughly half a year into the writing process. Such an intermittent solicitation, though beneficial to inform the development process, is not the norm. The vast majority of groups tasked with the development of AI Ethics Principles solicit feedback ex ante (if at all). Generally, little to no leeway is given for amendments during the writing process, although some other notable exceptions can be found in e.g. the Australian AI Ethics Framework (DISER, 2019) or the New Zealand Draft Algorithm Charter (New Zealand Government, 2020). With the AI HLEG, this dynamic procedure allowed for amendments of initial propositions and engaged stakeholders to co-create an ethical framework, taking their concerns, unique expertise and suggestions into consideration.

[91] See: https://ec.europa.eu/digital-single-market/en/european-ai-alliance.
[92] See: https://ec.europa.eu/transparency/regexpert/index.cfm?do=faq.faq&aide=2.

Subsequently, the AI HLEG presented revised Ethics Guidelines in April 2019. Shortly thereafter, the AI HLEG continued to work in an agile manner, stress-testing the suitability of their recommendations in the real world. More precisely, in June 2019, two months post-publication of their final Ethics Guidelines, the European Commission opened a 'piloting phase' on behalf of the AI HLEG. This phase concerned itself with the third section of the Ethics Guidelines, which contained an assessment list meant to support the actionability of the Ethics Guidelines' key requirements, i.e. the main recommendations. As such, this piloting phase was meant to trial the usefulness, comprehensiveness and suitability of this list. The goal was to receive feedback that would allow the AI HLEG to improve the assessment list encouraging better actionability (AI HLEG 2020). Feedback was solicited through a three-pronged approach: (i) via 50 in-depth day-long interviews with selected companies; (ii) two quantitative surveys for technical and non-technical stakeholders; and, (iii) a dedicated space on the AI Alliance where feedback could be submitted. Together, this allowed for a breadth of multidimensional input.

1.3. Toolboxes

The 'toolboxes' of mechanisms accompanying the Ethics Guidelines enabled operationalisability of the key requirements. These toolboxes contained both technical and non-technical methods and recommendations which could be used independently, simultaneously, or, consecutively. The technical toolbox proposed to make use of e.g. architectures for trustworthy AI, testing and validation methods, explainable AI (XAI) research, as well as Quality of Service indicators. The non-technical toolbox ranged from public awareness and diversity measures, to efforts that can be undertaken by governments such as standardization, certification, and regulation. The fact that the AI HLEG included non-technical methods demonstrates that the responsibility to co-create, maintain and deploy trustworthy AI extends well beyond the technical arena. Similarly, the range of

methods proposed reflected the range of stakeholders that are necessary to develop 'trustworthy AI': from governments with the mandate to create regulation, to researchers with their ability to shape the implications and features of their AI development process, and from industry actors who are in a position to create more diverse hiring processes (Crawford, 2016), to civil society actors with their power to demand and engage in multi-stakeholder dialogues.

In summary, the AI HLEG process highlighted the following procedural mechanisms: (i) involvement of diverse voices, both between the experts and through open public feedback; (ii) agile development process through interim and ex post feedback processes; and, (iii) mechanisms and methods to support actionability. These aspects appear particularly promising components for grounding and supporting an initial approach to developing Actionable Principles.

2. A preliminary framework for Actionable Principles

The above brief review of the development process of the Ethics Guidelines indicates three promising elements: support of diversity, allowance for agile development and support for implementation. This section builds on these elements, in order to generalize them and expands their scope. It proposes a preliminary framework towards Actionable Principles composed of the following: (1) preliminary landscape assessments; (2) multi-stakeholder participation and cross-sectoral feedback; and (3) mechanisms to support implementation and operationalizability.

These steps cover crucial turning points at each stage of the development process towards Actionable Principles, from inception to development, to their

post-publication stage. A (1) *preliminary landscape assessment* addresses the contextual environment within which Actionable Principles arise. Once that has been established, (2) *multi-stakeholder participation and cross-sectoral feedback* addresses the composition and working methods of those parties which draft any sets of Actionable Principles. Finally, (3) *support of implementation and operationalizability* addresses the direct move towards Actionable Principles' implementation into governmental policy-making post publication. In discussing each of the three elements of this prototype framework in turn, it can be seen how they also relate to - or provides insights on - some of the common critiques of existing AI Ethics Principles.

2.1. Development of preliminary landscape assessments

Actionable Principles benefit from what this paper calls a 'landscape assessment'. This could inform their development process, and serve to place Actionable Principles within the particular environment (geopolitical, societal, legal etc.) in which they are implemented, as well as to identify blindspots or practical difficulties before they arise. In a similar manner to government departments comparing various policy options in order to decide on the best one to implement, or to conduct impact assessments prior to introducing new regulation, landscape assessments can support foresight and analysis in the drafting of Actionable Principles, making them more impactful and applicable.

Landscape assessments therefore create a bridge between what should ideally be done, and what can be done (and how), supporting a step change into the right direction. Promising areas for a landscape assessment include the technical state-of-the-art, identifying which capabilities and applications exist, what societal, economic and political factors may affect their potential proliferation or market

penetration to various actors, and the resulting timelines of sociotechnical change; such assessments also include the societal environment, to determine what are the public's and policymaker's overall understanding, range of concerns, and ability to engage with issues. Finally, it could serve to review the legislative status quo, to understand the scope of issues already covered (Gaviria 2020).

Overall, landscape assessments are likely to concretize and strengthen an ethical principle such as 'an AI should not unduly influence human agency' which otherwise might be too broad (that is, not specific enough about its requirements), or too overarching in the scope of the AI applications it seeks to apply to in order for it to be actionable and functional in policy. For example, a landscape assessment could address the technical state-of-the-art for synthetic media, identifying a lack of technical tools to adequately capture all synthetic audio and visual products, and point towards civil society being insufficiently educated about this technological capability. A landscape assessment might also identify that the Californian B.O.T. Act (Bolstering Online Transparency Act, 2018) requires bots on the internet to self-identify in order to not mislead humans, however, highlight that this is not generally applicable to all 'output' derived from an AI agent.

Similarly, an assessment of the state-of-the-art of energy-efficient learning in AI, and tradeoffs between beneficial AI applications now versus the impact of climate change going forwards, would have value. For example, this could ensure that ethical considerations for future generations or tensions between developing technological solutions now at a potential longer-term cost can be clearly flagged and evaluated in advance. In short, a landscape assessment could support the identification of practical issues that map onto more theoretical aspects, facilitating more actionable policy. Actionable Principles informed by landscape assessments can serve to form the backbone for timely and evidence-based policy making. Moreover, they address a common criticism that AI Ethics Principles lack access to adequate information necessary to make impactful recommendations.

A lack of adequate and pertinent information, or the ability, resources and authority to gather such, underlying the development process of any set of AI Ethics Principles, can jeopardize the resulting impact significantly. In particular, it could hinder the ability to foresee practical second-order issues that may result out of the recommendations provided, or to make truly impactful and actionable recommendations.

Looking back over the past years, various cases illustrate how a lack of advanced landscape assessments can constrain the downstream impact of AI Ethics Principles. In 2017, New York City enacted bill "Int. 1696", establishing an Automated Decision Systems Task Force (ADS). The goal of this task force was to develop recommendations reviewing New York City's use of automated decision-making systems, including individual instances of "harm". However, their final report was subject to critique, which amongst other points homed in on its weak recommendations, which were partially traced back to the lack of resources the ADS had initially been allocated to conduct an appropriate landscape assessment. On this basis, AI Now Institute's Shadow Report on the ADS's work (Richardson, 2019) emphasized that any task force or group meant to "review, assess and make recommendations" should be empowered to receive all necessary information to the fulfillment of their task in order to be able to appropriately evaluate the landscape and provide informed outputs. This includes access to existing laws, policies and guidelines. Going forwards, stakeholder groups drafting Actionable Principles need to be able to ensure that the impact of their work is not minimized by a lack of, or, ability to conduct a landscape assessment.

A landscape assessment could address practical hurdles of actionability that result from a lack of access to sufficiently comprehensive and relevant information. It therefore constitutes the first pillar in formulating Actionable Principles for AI.

However, while necessary, it is not sufficient by itself to ensure actionability. For this, we must turn to additional mechanisms.

2.2. Multi-stakeholder participation and cross-sectoral feedback

In order to achieve an outcome that works for all, the stakeholder group developing Actionable Principles must be diverse and representative. This allows for a broad (and relevant) range of concerns and recommendations to be taken into account and accurately reflected. After all, one goal of Actionable Principles is to encourage good policy-making that works for all of society. Despite being a relatively diverse multi stakeholder group, the AI HLEG's composition could have been improved, e.g. when it came to the balance between civil society and industry representatives (Hidvegi & Leufer, 2019). Indeed, during the composition of multi-stakeholder groups, special care should be taken that those who are most likely to be adversely affected due to the development, deployment and application of certain AI systems are adequately represented, have their opinions heard, and can have an outsized say on the issue at hand.

Intra-group diversity is necessary, it is not sufficient. Instead of focusing on the ideal building blocks for intra-group diversity, this paper will explore an expansion to this goal. For Actionable Principles this could take the form of significant agile engagement with multiple cross-sectoral stakeholders outside of the group, consulting them on their expertise, insights and unique point of view. Elsewhere, such an approach has also been proposed for 'human rights-centered design, deliberation and oversight' of AI (Yeung et al., 2019). It ensures that a much larger diversity of voices is captured that otherwise would be missing in the process, regardless of the composition or size of the group. This is an important consideration for the development of Actionable Principles because no group, no

matter how diverse, can possibly reflect the knowledge and expertise of hundreds of diverse stakeholders. Moreover, shortcomings such as group think are more likely when groups are unable to actively engage with outside stakeholders.

The predominant method of soliciting external expertise is to conduct a public consultation. Broadly speaking, there appear to be three different approaches: (i) consultations prior to the drafting stage; (ii) consultations during the drafting stage; and, (iii) consultations post publication stage.

Most commonly, (i) consultations are conducted prior to the drafting stage. These consultations often take the form of either presentations to the stakeholder group tasked with the drafting, or the submission of written position papers. These submissions are often based on a set of questions meant to inform and guide their content. For example, the report of the Select Committee on AI from the House of Lords (House of Lords, 2018) was written based on feedback received in this manner.

While not replacing a comprehensive landscape assessment, an ex ante consultation can act to inform of the concerns, suggestions and existing difficulties that multiple stakeholders experience or foresee in the nearer future on a range of topics. Indeed, in the context of broader technology policymaking, the paradigm proposed by the Responsible Research and Innovation Framework has long emphasized such consultations (Owen et al., 2013; Stilgoe et al., 2013), suggesting they should play a crucial role in structural frameworks for the development of Actionable Principles. At the same time, ex ante consultations, particularly without a preliminary landscape assessment to guide them, suffer from shortcomings that could negate their usefulness. In particular, they often shift the burden of knowing what could constitute 'useful' information on the person or entity submitting information. They equally force the stakeholder group soliciting that information to already have a clear sense of the type of information they need, by virtue of asking the 'correct'

questions or call for input. This can lead to two issues that constitute reverse sides of the same coin: those submitting may submit overly generic feedback, and the stakeholder group may ask overly broad questions. Both ultimately lead to types of feedback that may end up lacking (or swamping) the information that would have been relevant in hindsight.

A consultation prior to the drafting stage would therefore benefit from building on a preliminary landscape assessment. This would increase the clarity as to what type of additional information is sought from stakeholders and ensure that time on submissions is well spent, especially bearing in mind that some groups may not be able to spend significant amounts of time on submissions. This type of consultation would also benefit from being supplemented by the second and third type of consultation — an (ii) intermittent consultation during the drafting stage and a (iii) consultation post-publication stage, ensuring a flexible development process and timely feedback loop. Unfortunately, both (ii) and (iii) are rarely if ever made use of. The following paragraphs will outline why Actionable Principles would, however, benefit from these tools.

An (ii) intermittent consultation during the drafting stage can serve to stress test an initial proposal and steer the work towards actionability, in accordance with valid critiques and additional information derived from the consultation. Such a dynamic feedback loop can serve to elevate work from expert led to multi-stakeholder led with crowdsourced input. In addition to such an intermittent consultation on the Ethics Guidelines, stakeholders had the opportunity to engage directly with the AI HLEG via the European AI Alliance. Elements such as these could be useful to increase the comprehensiveness of Actionable Principles.

The third avenue of consultation, (iii) post publication stage, likewise remains missing from most published AI Ethics Principles. Such a consultation could assist in understanding reasons for a potential lack of implementation, or pinpoint

omissions, outdated assumptions, or areas that require revision since the moment of publication. This type of consultation could support either a revision of the outcome or serve as a guiding structure for future attempts. For the purpose of achieving Actionable Principles, the former would be preferable as it has the potential to refine and adapt the document over time, strengthening it and encouraging the interplay between ethics and agile policy making.

All three avenues for consultation support the inclusion of different types of expertises, cross-sectoral dialogues and understandings, while counteracting a potential 'ivory tower mentality' which a stakeholder group closed to engagement with those external to the group may easily fall prey to.

One angle of concern, or criticism, is whether existing AI Ethics Principles accurately capture public and expert led-consensus. It should be noted that on many topics, consensus may be impossible to reach despite broad stakeholder inclusion. Green (2018) argues that there are certain topics which may wrongly be perceived as a constant, when they are actually a dynamic social construct. Take, for example, 'safety'. In an analysis by Fjeld et al. (2020) of AI Ethics Principle-type documents, over 81% contained some recommendations in favor of 'safety' and 'security'. Despite this apparent high-level commonality and agreement, they all suffer from a major deficit in light of Green's (2018) argument. Specifically, he argues that perceptions of what constitutes safety thresholds differ amongst people and societies. This would entail that e.g. AI would need to pass the median between all individual safety thresholds distributed across society, in order to be perceived as safe even if the concept is understood by all to denote the same thing. Whilst some ethical concerns may intuitively be perceived to have 'fuzzy boundaries', few would say the same about a concept such as 'safety'. Moreover, Jobin et al. (2019) identify significant variations in the different interpretations of commonly shared principles such as transparency. The likelihood of 'fuzzy boundaries' embedded in recommendations of Actionable Principles, be that on safety, environmental

wellbeing, transparency or others, demonstrates the overall importance of introducing a multi-stakeholder approach reflecting as many considerations and concerns as possible. This can support Actionable Principles to avoid baking in 'suitcase terms' that lend themselves to diverse interpretation and future contestation.

Building on the aforementioned point, various groups, depending on their background, can end up using the same terminology, yet denoting different content spaces resulting in varying interpretations for their resolution. This creates the appearance of agreement at the drafting stage while causing difficulties at the implementation stage. For example, Xiang and Raji (2019) analyzed various ways in which the technical research community makes use of legal terminology around fairness, but often lacks the necessary legal understanding to align their interpretation with what fairness is assumed to mean from a legal point of view, and vice versa. This misalignment between terminology and interpretation can quickly become a practical policy problem. Moreover, it can yield methodologically different strategies to address what is perceived as the same recommendation at hand (Whittlestone et al., 2019). The same recommendation may be seen as a political, ethical or technical problem, depending on the stakeholders involved trying to resolve it. Finally, this may introduce questions whether or not data-driven or algorithmic approaches should be outright banned in certain situations where individuals stand to be harmed by a subpar solution, or where the technical capability is not set to meet the requirements necessary from a societal and legal point of view (Wachter et al., 2020).[93]

In conclusion, in order to support Actionable Principles ongoing multi-stakeholder debate and broad continuous information exchange via e.g. multiple public consultations is key.

[93] This section does not aim to resolve whether or not fairness can be automated but how stakeholder dialogue could serve to identify tensions between different stakeholders' opinions and approaches.

2.3. Mechanisms to support implementation and operationalizability

Most AI Ethics Principles are drafted without concrete plans for their implementation, impact or target audience (Schiff et al., 2019). Indeed, they are often abstract and make vague suggestions (Mittelstadt, 2019). This can hinder their actionability and implementation in policy-making. Guidance in the form of a toolbox, or method to operationalize the recommendations can be a crucial step to move from AI Ethics Principles towards Actionable Principles (Morley et al., 2019). The benefit of providing existing and desired tools is equally taken up by proponents of a human rights frameworks based approach with the expectation that these can be adapted to ensure an integration of human rights norms within the AI lifecycle (Yeung et al., 2019). Accompanying methods and measures stand to strengthen Actionable Principles by providing guidance as to how their recommendations should be implemented and by whom. Moreover, recent academic work on mechanisms for supporting verifiable claims (Brundage et al., 2020) highlighted the importance of providing such mechanisms to actors involved throughout the AI lifecycle, including governments.

In order to account for the broad impact of AI systems it is necessary to consider and provide both technical and non-technical tools and measures. Indeed, the Ethics Guidelines did both. Technical measures could for example cover explainable AI research (XAI), which can help to illuminate some of the underlying decision making processes within some AI systems, increasing an individual's ability to understand (and challenge) the AI system's output. They could also entail privacy-preserving measures, which serve to adequately protect and secure (personal) data used to develop and maintain a given AI system's functionality; requirements for sufficiently representative data sets (Flournoy et al., 2020) to ensure both adequate functioning of an AI system in a real-world environment as well as to tackle bias introduced through inadequate data sets; or, describe testing and validation protocols and procedures that would support the fulfillment of the

principles in question by a given AI system. The latter may cover adversarial testing by dedicated red teams (Brundage et al., 2020), evaluation of out of distribution robustness (Lohn, 2020), repeat testing within reasonable time spans once deployed, or denoting adequate requirements for audit trails at all stages of the lifecycle of a given AI system (AI HLEG, 2020).

Non-technical (or less technical) measures can provide requirements and methods to increase civil society's participation in decision-making processes surrounding the development and deployment of AI systems, empowering them to (safely) call out and halt the development and application of AI systems with potentially negative impacts. The dialogue with civil society and affected parties could be strengthened through e.g. the development and provision of relevant 'algorithmic impact assessments' (Reisman et al., 2018). An example of this could be the Canadian government's algorithmic impact assessment[94] or the AI HLEG's revised assessment list for trustworthy AI (AI HLEG 2020). Other measures could cover hiring practices, institutional mechanisms whereby employees can safely flag concerns surrounding a given AI system (e.g. its performance or scope), or the creation of audit trails (e.g. to enable the auditing of AI systems by third parties) (Raji et al., 2020). Moreover, other relevant organizational efforts such as incentivizing AI developers through codes of ethics or codes of conduct could be suggested.

Finally, as stated earlier, Actionable Principles may influence policy-making simply by virtue of shaping governmental funding decisions. Along these lines, principles that concern e.g. the protection of the environment and a support for the flourishing of future generations (AI HLEG, 2019b) could benefit from being accompanied by recommendations for funding research into methods such as computationally efficient algorithms (Strubell et al., 2019).

[94] See: https://open.canada.ca/aia-eia-js/?lang=en.

While the provision of mechanisms for implementation does not replace a clear and coherent structure of the document in question, it can contribute to an actionability of underlying goals. Furthermore, this directly builds and expands on the other two elements: the landscape assessment and multi-stakeholder engagement strategy. Together forming a prototype framework across the lifecycle of Actionable Principles — inception, development, post-publication.

As discussed earlier, a prominent critique of AI Ethics Principles at large is that they are susceptible to 'ethics washing', including concerns that they may run risk to morph away from moral principles into a 'performative facade' (Bietti, 2019) and, ultimately, that they are an 'easy' or 'soft' option (Wagner, 2018) in comparison to 'hard(er)' governance mechanisms such as regulation. This paints them as a distraction rather than a solution to the problems they are hoping to address. In fact, even members of AI HLEG have criticized the Ethics Guidelines over 'ethics washing', alleging excessive industry influence (Metzinger, 2019). It should be noted that these criticisms are largely launched against the role of industry in operationalizing AI Ethics (Ochigame, 2019) or developing AI Ethics Principles in order to forestall or evade government regulation and real oversight. The focus of this paper, however, is on governments and their policy-making, i.e. on *increasing* the influence Actionable Principles can have on concrete soft and hard governance developments. It sees ethical enquiry as key to developing good policy even if the focus of the framework is currently limited to procedural aspects as enablers.

In conclusion, it is suggested that the provision of methods and mechanisms to increase the actionability of recommendations in Actionable Principles through e.g. testing and validation protocols, cross-societal dialogue and audit trails further closes the gap by providing tools to move from principles to policy.

3. Conclusion

While their rise has been encouraging and needed, the majority of AI Ethics Principles today still suffer from a lack of actionability in policy-making. This paper has suggested that instead of abandoning the approach altogether, the community should iterate on and improve these tools in order to produce a form of 'Actionable Principles' on AI, which integrate actionability and ethical reasoning. In order to do so, this paper first briefly identified and examined a series of promising elements from the development process of the Ethics Guidelines by the AI HLEG. Subsequently, it proposed three elements towards a framework for Actionable Principles: (1) preliminary landscape assessments; (2) multi-stakeholder participation and cross-sectoral feedback; and, (3) mechanisms to support implementation and operationalizability.

Given the current pace of governmental initiatives and the drastically increasing number of groups working to develop relevant guidance, this paper's suggestion comes at a crucial time. Excitement about AI Ethics Principles has waned as the reality has set in that significant work from all stakeholders needs to be undertaken to move them from paper to practice. Multiple efforts are underway to do so (e.g. via the Global Partnership on AI's committee on responsible AI[95] or the German government's project on 'ethics of digitalization'[96]) and it is hoped that the prototype framework provided in this paper sets out a useful path. Inevitably, actionability is not and will not be the only hurdle AI Ethics Principles will face over the coming years. Nevertheless, the pacing-problem (Marchant et al., 2011) is real and future governance efforts will significantly rely on existing (academic) work as they make sense of direly needed policy options. This puts Actionable Principles in a potentially powerful political and societal role, one that needs to be taken seriously and nourished.

[95] See: https://oecd.ai/wonk/an-introduction-to-the-global-partnership-on-ais-work-on-responsible-ai
[96] See: https://www.bundespraesident.de/SharedDocs/Berichte/DE/Frank-Walter-Steinmeier/2020/08/200817-Ethik-der-Digitalisierung.html

Bibliography

AI HLEG. (2019). A Definition of AI: Main Capabilities and Disciplines. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. European Commission, High Level Expert Group on AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

AI HLEG. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Article 19. (2019) Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence. https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

BEUC. (2019). AI Ethics Guidance: a first step but needs to be transformed into tangible rights for people. https://www.beuc.eu/publications/beuc-pr-2019-011_ai_ethic_guidance.pdf.

Bietti, E. (2019). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. *Proceedings of ACM FAT* Conference (FAT* 2020). ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/3351095.3372860.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H. et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. http://arxiv.org/abs/2004.07213.

Bolstering Online Transparency Act, Senate Bill No. 1001 (2018) (enacted).

https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

Crawford, K. (2016). Artificial Intelligence's White Guy Problem. *The New York Times*. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Defence Innovation Board (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

DISER. (2019). Department of Industry, Science, Energy and Resources, Australian Government. AI Ethics Framework. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework.

Donahoe, E., Metzger, M. (2019). Artificial Intelligence and Human Rights. *Journal of Democracy*, Volume 30, Number 2, April 2019, pp. 115-126. Johns Hopkins University Press.

European Commission. (2018). Communication Artificial Intelligence for Europe. COM(2018) 237 final. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

European Commission. (2020). White Paper on Artificial Intelligence - A European Approach to Excellence and Trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Gaviria, C. (2020). The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the U.S.https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html.

Klöver, C., Fanta, A. (2019). No red lines: Industry defuses ethics guidelines for artificial intelligence.

https://algorithm.watch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/.

Hidvegi, F, Leufer, D. (2019). Laying down the law on AI: ethics done, now the EU must focus on human rights. https://www.accessnow.org/laying-down-the-law-on-ai-ethics-done-now-the-eu-must-focus-on-human-rights/.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. https://doi.org/10.2139/ssrn.3518482.

Floridi, L., Cowls, J., Beltrametti, M. et al. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* 28, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5.

Flournoy, M., Haines, A., Chefitz, G. (2020). Building Trust Through Testing. https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8.

House of Lords. (2018). AI in the UK: ready, willing and able?. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Lohn, A. (2020). Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance. http://arxiv.org/2009.00802.

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem. *Springer Science & Business Media*.

Metzinger, T. (2019). EU guidelines: Ethics washing made in Europe. Der

Tagesspiegel. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Mittelstadt, B. (2019). AI Ethics -- Too Principled to Fail?. http://arxiv.org/abs/1906.06668.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00165-5.

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Vol. Summer 2020, pp. 1-70). Palo Alto: CSLI, Stanford University. https://plato.stanford.edu/entries/ethics-ai/.

New Zealand Government. (2020). Draft Algorithm Charter. https://data.govt.nz/use-data/analyse-data/government-algorithm-transparency-and-accountability/draft-algorithm-charter/.

Ochigame, R. (2019). The invention of "Ethical AI" how big tech manipulates academia to avoid regulation. *The Intercept*. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/?comments=1.

OECD. (2019). OECD Principles on AI. https://www.oecd.org/going-digital/ai/principles/

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A framework for responsible innovation. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, *31*, 27–50.

Raji, D., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. DOI:https://doi.org/10.1145/3351095.3372873.

Reisman D., Schultz J., Crawford, K., & Whittaker, M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, *AI Now*, April 2018.

Rességuier, A., Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, Vol. 7(2), 2020. 85.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*.

Richardson, R., ed.. (2019). Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. *AI Now Institute*. https://ainowinstitute.org/ads-shadowreport-2019.html.

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2019). What's Next for AI Ethics, Policy, and Governance? A Global Overview. https://doi.org/10.31235/osf.io/8jaz4.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. http://arxiv.org/abs/1906.02243.

Veale, M. (2020). A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence. *European Journal of Risk Regulation*. https://ssrn.com/abstract=3475449.

Von der Leyen, U. (2019). A Union That Strives for More. My Agenda for Europe.https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf.

Voss, A. (2020). JURI Draft report: Civil liability regime for artificial intelligence. PE650.556. European Parliament. https://www.europarl.europa.eu/doceo/document/JURI-PR-650556_EN.html?redirect.

Wachter, S., Mittelstadt, B., Russell, C. (2020). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI.

http://dx.doi.org/10.2139/ssrn.3547922.

Whittlestone, J., Nyrup, R., & Alexandrova, A. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield*. http://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf.

Whittlestone, J., Ovadya, A. (2019). The tension between openness and prudence in AI research. http://arxiv.org/abs/1910.01170.

Xiang, A., Raji, I. D. (2019). On the Legal Compatibility of Fairness Definitions. http://arxiv.org/abs/1912.00761.

Yeung, K., Howes, A., Pogrebna, G. (2019). AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing. M Dubber and F Pasquale (eds.) *The Oxford Handbook of AI Ethics, Oxford University Press.* https://ssrn.com/abstract=3435011.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles. http://arxiv.org/abs/1812.04814.

Chapter IV

The case for an 'incompletely theorized agreement' on artificial intelligence policy

……..

Stix, C., Maas, M.M. Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy. *AI Ethics* 1, 261–271 (2021). https://doi.org/10.1007/s43681-020-00037-w.

……..

Introduction

The prevailing uncertainty around the trajectory and impact of artificial intelligence (AI) makes it clear that appropriate technology policy matters today to ensure that AI enables broad societal benefit. The possible ethical and legal impacts are vast: from algorithmic bias to AI-enabled surveillance, and from lethal autonomous weapons systems to possible widespread technology-induced unemployment. Moreover, some forecast that continuing progress in AI capabilities will eventually make AI systems a 'general-purpose technology' [1], or may even in time enable the development of forms of 'high-level machine intelligence' (HLMI) [2] or other 'transformative AI' capabilities [3]. Debate on these latter scenarios has been diverse, and has at times focused on what some have referred to as 'Artificial General Intelligence' (AGI) [4]. On the surface, those concerned with AI's impact can appear divided between scholars who focus on discernible problems in the near-term, and scholars who focus on less certain problems in the longer-term [5–8].

This paper wants to investigate and critically examine the dynamics and debates between these two communities, and prospects for convergence or functional policy

cooperation. As such, the focus is not on the relative plausibility of various advanced AI scenarios such as HLMI or AGI, nor do we mean to suggest that a long-term perspective is solely focused or concerned with AGI [9–11]. Rather, we pragmatically examine the current dynamics and organization of the growing AI community, with an eye to policy effectiveness.[97] This paper proposes that insofar as a divergence exists, it is not productive to the technology policy goals of either group within the community with regard to shaping responsible and ethical AI policy.[98] It suggests that differences may be overstated and proposes that even if one assumes deep differences, these are not practically insurmountable. Specifically, it argues that the constitutional law principle of an 'incompletely theorized agreement' provides both theoretical foundations and historic precedent for cooperation between divergent communities or actors without compromising respective goals. In doing so, this paper seeks to take up the recent call to 'bridge the near- and long-term challenges of AI' [8].

It proceeds as follows: in Section 2, we briefly lay out the landscape of AI policy concerns and the structure of the associated AI ethics and policy community. In Section 3, we explore three potential sources which could contribute to divergence within said AI ethics and policy community: (a) *epistemic,* relating to different thresholds for uncertainty; (b) *normative,* relating to different perceptions of value; (c) *pragmatic,* relating to the tractability of formulating AI (policy) actions today that maintain long-term relevance. Finally, in Section 4, we propose that one consolidating avenue to harness mutually beneficial cooperation could be anchored in the constitutional law principle of an 'incompletely theorized agreement'. This

[97] We do not seek to take positions over the direct, first-order questions such as the (im)plausibility of future advanced AI or its timelines, nor on the relative ethical importance of 'long-term' vs. 'near-term' concerns.

[98] In this context, we define 'AI policy' as concrete soft or hard governance measures which may take a range of forms such as principles, codes of conduct, standards, innovation and economic policy or legislative approaches, along with underlying research agendas, to shape AI in a responsible, ethical and robust manner. Our paper works under the assumption that policy making can positively influence the development and deployment of AI.

argument works with the assumption that governmental policy making is better influenced through collaboration as a united community than fragmented smaller attempts by subgroups.

2. AI policy: A house divided?

Recent progress in AI development and deployment has given rise to an array of ethical and societal concerns.[99] Accordingly, there have been calls for appropriate policy measures to address these, and to ensure that AI is developed and deployed to the benefit of society.

As an "omni-use technology" [12] AI yields potential for good [13–15] and for bad [16–18]. The latter include: various forms of pervasive algorithmic bias [19,20], challenges around transparency and explainability [21,22]; the safety of autonomous vehicles and other cyber-physical systems [23], or the potential of AI systems to be used in (or be susceptible to) malicious or criminal attacks [24–26].

Moreover, societies may have to reckon with privacy concerns in the face of widespread surveillance [27]; with the economic and political effects of technological unemployment [28,29]; the erosion of democracy in the face of tech companies' corporate power, 'computational propaganda' or 'deep fakes' [30–32], or an array of threats to various human rights [33,34]. Some have even anticipated the possible erosion of the global legal order by the comparative empowerment of authoritarian states [35,36]. Finally, some express concern that continued technological progress might eventually result in increasingly more 'transformative' AI capabilities [3], up to and including AGI. Indeed, a number of AI researchers expect some variation of 'high-level machine intelligence' to be achieved within the next 5 decades [2]. Some have suggested that, if those transformative capabilities are not handled with

[99] This paper perceives ethical and societal concerns to be closely intertwined, and refers to the broader set of these actual and potential concerns throughout.

responsibility and care, such developments could well result in new and potential catastrophic risks to the welfare, autonomy, or even survival of societies [37,38].

Looking at the current debate and scholarship involved in the aforementioned areas, we note, along with other scholars [5,7,39], that there appears to be a temporal split, along a 'near-term'/'long-term' axis. This is not a distinction that turns on specific timeframes (e.g. 'within 10 years' vs. 'beyond 50 years'). Nonetheless, Baum has broadly characterized a distinction between 'presentist-' and 'futurist factions', whose core claims, he takes to be that, respectively, "[a]ttention should go to existing and near-term AI", and "[a]ttention should go to the potential for radically transformative long-term AI" [5]. Similarly, Cave and ÓhÉigeartaigh characterize the former 'faction' as focusing on "immediate or imminent challenges involving fairly clear players and parameters, such as privacy, accountability, algorithmic bias and the safety of systems that are close to deployment", while the latter focuses on "longer-term concerns and opportunities that are less certain, such as wide-scale loss of jobs, risks of AI developing broad superhuman capabilities that could put it beyond our control." [8].

A perceived or experienced distinction like this may become a self-fulfilling prophecy that contributes to a real separation over time. This can be the case even if the perceived differences are based on misperceptions or undue simplification by popular media, as may have happened to the field of nanotechnology [40]. This includes (but is not limited to) some public interaction between groups which have appeared strained. From the near-term perspective, 'long-term' concerns have at times been dismissed as "science fiction at best or an irresponsible distraction at worst" [41]. Here, the underlying worry seems to be that such problems are speculative or at least many decades away [42] and distract from other urgent problems such as the direct ethical, economic, political or geostrategic effects of current AI systems [16,43,44]. At the same time, it has been noted that much of the media coverage and public debate fails to engage with long-term points of concern

[45,46], instead invoking caricatured notions of 'the singularity', or are colored by misleading popular-cultural depictions of anthropomorphic 'rebellion' by 'malevolent' robots [47].

From the long-term perspective, although the validity of many short term concerns is acknowledged, it is sometimes argued that their importance must be set against the potential extreme negative risks hypothetical future AI advances could produce [37]. At the same time, one does not even need to accept the possibility of HLMI to accept that near-term capabilities can be (indeed, arguably have already been) transformative at a global level, even up to the point of posing catastrophic risks [9,48]. It is important to note that there is a remarkable amount of outstanding variety and points of view in the general community concerned with the impacts of AI. This is equally the case among those focusing on longer-term risk where, many scholars remain divided over when, if ever, machines might reach general human-level performance [2]; how rapid or sudden this progress might be [49,50]; what shape such AI systems could or would take [10,50]; and whether or not such a development, if possible, would necessarily pose a risk [51,52]. It is frequently emphasized that there is extensive difficulty in accurately predicting future progress and breakthroughs in AI development [53], with many past predictions being wildly unreliable [54].

Whatever the positions in question, recent years have seen unproductive engagement between scholars and spokespersons that identify with either community. This provides an interesting occasion to investigate the possible origins, effects, and necessity of this alleged distinction.

3. Examining potential grounds for a division: epistemic, normative, pragmatic

Accordingly, the following mapping provides an early attempt at a taxonomy of possible sources that could account for a clustering into fuzzy 'near-' and 'long-term' communities. It is not the result of a comprehensive opinion survey. In our taxonomy, we suggest three potential sources; (a) epistemic; (b) normative; and (c) pragmatic. They are not mutually exclusive, and different scholars may hold distinct and overlapping sets of beliefs on any of the following axes [cf. 7].

3.1. Epistemic distinctions

In this case, *epistemic* differences may reflect varying levels of tolerance regarding scientific uncertainty and distinct views on the threshold of probability required before (far-reaching) action or further investigation is warranted. In other words, from the perspective of expected value theory, some might argue that the extremely high stakes involved in some advanced AI scenarios would warrant that these matter merits *some* investigation, even if the probability is very low (or unknown) [55], because oven small reductions in the probability of a potential risk would have enormous value [56]. Conversely, scholars focused on shorter-term concerns may argue that there are too many 'unknown unknowns' around future scenarios involving HLMI or AGI systems for any meaningful quantifiable estimate of probability to be made. Epistemically, an argument could be made that the demonstrated and measurable risks that AI yields today (or risks that are uncertain but meet some minimum standard of plausibility) merits our attention instead. This could include a worry that if we were to accept an expected-value heuristic entirely unshielded by this minimum threshold of 'acceptable probability', then this would force us into the perverse situation where we are compelled to accept as serious problems, many speculative scenarios that appear highly implausible, while abandoning or ignoring many clear and present dangers today. All this goes to say that the topics of concern for various AI policy issues may depend on *qualitatively*

different conceptions of 'acceptable uncertainty'. Different judgments on this conception may well be hard to resolve.

Moreover, *epistemic* differences might turn on implicit or explicit disagreements over the modal standards that should apply in debates around longer-term and far-future concerns with regard to AI. That is, it turns on debates over what types of data or arguments are considered admissible as evidence in establishing or contesting the plausibility or probability of risk from advanced AI systems. Should philosophical argumentation be admitted, and how should this be set against technical arguments? It also turns on different views over the *validity of modelling assumptions* in forecasts: can these introduce significant scientific breakthroughs, or must they be restricted to extrapolation from present-day technological trajectories (e.g. Moore's law)? Finally, they turn on *differential interpretations of evidence* that is available. For instance, do empirically observed failure modes of present-day architectures [57–60] provide small-scale proof-of-concepts of the difficulties we may encounter in AI 'value alignment', or is such an extrapolation invalid?

Indeed, these questions reflect complex, deep philosophical precommitments--about the nature of (scientific) knowledge, and about legitimate heuristics for making (precautionary) decisions in the context of uncertainty.

3.2. Normative distinctions

Some divergences may derive from a potential perception that, object-level issues aside, different community members weigh their shared core values differently. Overall, prominent values in debates around AI--in both academia and in public discourse--often appeal to widely shared ethical principles such as fairness, accountability, the rule of law, human rights, equity, social justice, and democratic principles [31,61], and, especially in the context of military applications of AI, values such as peace, security, stability and the international rule of law. In general,

these core values are relevant to all, with concerns over future advanced AI often motivated by an ethical perspective that cares about securing values such as fairness and human welfare 'into the far future' [5]. The divergence then appears to be a result of the temporal focus of these values and where they are expected to have greater relevance.

However, such a possible distinction ought not become a misdirected stereotype. There is considerable heterogeneity in the normative motivations within either community--it is very possible and valuable to work on nominally 'near-term' AI issues (such as bias, or surveillance capitalism) exactly out of a concern of how these will set the trajectory of human society and well-being into the longer term. Moreover, insofar as these underlying normative differences exist, they are often not practically relevant, since in many contexts both communities cash them out in the same direction--that is, both communities have emphasized policies or strategies that promote broad societal inclusion of benefits to AI, whether this is out of a concern over injustice and inequality, or out of a consideration to ensure 'non-turbulence' and 'magnanimity' as important principles for a historically tumultuous and precarious societal transition towards advanced AI [62].

3.3. Pragmatic distinctions

Imperfectly described, *pragmatic* perceptions of the empirical dynamics and path-dependencies of policy or technical research aimed at pursuing responsible AI might contribute to divergent foci and actions across the community. This may lend itself to distinct 'theory-of-change', regarding the tractability and relevance of formulating useful and resilient policy action (or technical research) today. In this context, Whittlestone & Prunkl distinguish at least separate four questions on which communities appear to differ [7]: *Capabilities* (whether to focus on the impacts of current AI systems or those relating to advanced AI systems); *Impacts* (whether to focus on the immediate impacts of AI, or those possible impacts much further into the future); *Certainty* (whether to focus in impacts that are certain and

understood, or those that are more uncertain and speculative), or *Extremity* (whether to focus on impacts at all scales, or prioritizing those that could be especially large) [7]. In their reading, these four questions are distinct, and often confused or conflated into a binary distinction between 'near-term' and 'long-term', when a more accurate picture and more productive dialogue could be achieved by differentiating between these four dimensions of capabilities, impacts, certainty and extremity, and emphasizing that all of these sit on a spectrum [7].

Taking this point on board, there are additional ways to cash out possible differences. One debate might concern the question, *how long-lasting are the consequences of near-term AI issues?* If those that care about the long-term are convinced that these issues will not have long-lasting consequences, or that they would eventually be swamped by the much larger trends and much larger issues introduced by 'transformative' or otherwise advanced AI systems [3], then this could lead them to discount work on near-term AI problems. However, it is important to note that near-term issues--and the degree to which they are (mis)handled--are likely to considerably affect or frame the degree to which society is vulnerable or exposed to long-term dangers posed by future advanced AI systems. Short-term or medium-term issues [6,39] can easily increase society's general 'turbulence' [62], or lock in counterproductive framings of AI or our relation to it. In general, we might expect many nominally 'near-term' effects of AI on society (such as in surveillance; job automation; military capabilities) to scale up and become more disruptive as AI capabilities gradually increase [17,39]. Indeed, recently some long-term scholars have argued that advanced AI capabilities considerably below the level of HLMI might suffice to achieve 'prepotence' [9]. This would make such mid-term impacts particularly important to handle and collaboration between different groups especially fruitful.

Another pragmatic question or concern is over how much leverage we have today to meaningfully shape policies (or technical research agendas) that will be applicable

or relevant in the long term, especially if AI architectures and the broader political and regulatory environment changes a lot in the meantime [7]. Some scholars in the near-term community may hold that future AI systems, whatever their form, will either be so different from today's AI architectures that research into this question undertaken today will not be relevant; or conversely, they argue that such advanced AI capabilities might be so remote that the technological- and regulatory environments will both change too much for meaningful work to be conducted right now [63]. In conclusion, people in this position would argue that we had better wait until things are clearer and we are in a better position to understand whether and what research is needed or meaningful.

In practice, this critique does not appear to be a very common position. And indeed, as a trade-off it may be overstated. It is plausible that there are diverse areas on which both communities can undertake valuable research today, because the 'shelf life' of current policy and technical research efforts might be longer than is assumed. To be sure, there is still significant uncertainty over whether current AI approaches can be 'scaled up' to very advanced performance [64–66]. Nonetheless, technical research could certainly depart from a range of areas of overlap [67] and shared areas of concern [68,69].

Moreover, policy-making is informed by a variety of aspects which range across different time spans. Starting with political agendas that often reflect the current status quo, policy making is equally shaped by shifting public discourse, societal dynamics and impactful events. In the case of the latter and with regard to AI, this has often been negative events rife with discrimination, lack of transparency and threat to such as shifts in policy following the identification of discriminatory algorithms in UK visa selection processes [70], the Artificial Intelligence Video Interview Act regulating the use of AI in employee interviews, or the California B.O.T. Law requiring AI systems to self-identify as such [71,72]. Both near- and long-term communities share an interest in studying the dynamics of how *policy*

around AI is (and can be) formed by 'epistemic communities' [73], and what are the steps in the 'policymaking process' that determine which issues get raised to political agendas and eventually acted upon [74]. Given the above, research into the underlying social and societal developments is fruitful to develop and forecast mutually agreeable policy goals, across the 'policy making cycle' [75]. Insights into when and how AI research labs or individual researchers adopt--or alternately cut corners on--responsible and accountable AI, or how to incentivize shifts in workplace culture or employee norms to achieve this, can influence those developments and dynamics. Equally, the following might be suitable areas for combined research: research into the efficacy of 'publicly naming' biased performance results in commercial AI products on whether tech companies correct the biases in these systems [76], or research into whether 'codes of ethics' are actually effective at changing programmer decisionmaking on the working floor [77,78] could be of interest to either community, as are organizational proposals and frameworks for end-to-end organizational auditing of AI system development and deployment [79]. The question of how to promote prosocial norms in AI research environments is similarly of core interest to both communities with an eye to technology policy [80].

Both communities also share an interest in exploring debates over the *appropriate culture of disclosure or openness* [81,82], where it concerns AI applications with potentially salient misuses--a question of key relevance whether the concerns are over abuses of vulnerable populations, new vectors for criminal exploitation or attacks, new language models [83], or risks around future potent systems [84]. The same holds to some degree for research into *public perceptions* of AI (and their principals), and how these may be shaped by framings and narratives around the technology [17,85].

Finally, from a legal and policy perspective, it is important to note that while current regulatory work might not always prove directly transferable, in many

circumstances, it can provide the 'second-best' tools to use rapidly (and then adapt) to AI challenges, rather than wait out the slow and reactive formulation of policies. Significantly, there are usually important path-dependencies in technology regulation, such that AI initiatives today will likely have flow-through effects.

The points touched upon suggest that the strategic barriers to inter-community cooperation are not all that strong as is often made out--and that many 'tradeoffs' are overemphasized. But does that mean there are also positive, mutually productive opportunities for both communities to work on? And if such common opportunities exist, how could we ground collaboration between the communities--in spite of outstanding empirical or normative disagreements? In other terms, what would an incompletely theorized agreement for the AI ethics & policy community look like?

4. Towards an 'incompletely theorized agreement' for AI policy

We proposed some possible sources that could explain the potential split within the community, and why these may be largely unwarranted. This is especially the case when it comes to the development of suitable policy-making, benefitting from a united front with an eye towards societal well-being (be that short or longer-term). Accordingly, we will now discuss how even in the context of some prevailing epistemic or normative disagreements, these communities could reach agreement on shared policy goals and norms that remain in service to their (distinct) values and judgments.

Legal scholarship in constitutional law and regulation has long theorized the legal, organizational and societal importance of so-called 'incompletely theorized agreements' [86, 87]. These are agreements on practical outcomes to be achieved, reached amidst the deepest and most intractable theoretical disagreements. Incompletely theorized agreements are a fundamental component to

well-functioning legal systems, societies, and organizations that want to get urgent things done, against a backdrop of theoretical disagreement, showing a form of mutual respect [88]. Incompletely theorized agreements have distinctive social uses to maintain both stability and flexibility over time in the context of uncertainty (whether over principles or problems) [86]. They have long played a key role not just in constitutional and administrative law, but more broadly in making possible numerous landmark achievements of global governance, such as the establishment of the Universal Declaration of Human Rights.[100] Indeed, the incompletely theorized agreements framework has been extended to other domains, such as the collective development and analysis of health-care policies in the face of pluralism and conflicting views [92].

Incompletely theorized agreements also have broader similarities with the notion of an 'overlapping consensus', developed by John Rawls. This notion referred to the way that adherents of different (and apparently inconsistent) normative doctrines might nonetheless converge on particular principles of justice to underwrite the shared political community [93]. This concept has been read as one key mechanism in fields of bioethics, enabling agreement of members on an ethics committee or commission even under different comprehensive and fundamental outlooks [94]. Significantly, Seán ÓhÉigeartaigh and others have recently proposed such 'overlapping consensus' as one mechanism on which to ground *cross-cultural cooperation* on AI policy [91]. They point out that this concept has already played a role in the existing literature on computer ethics [95], as well as in the field of 'intercultural information ethics' [96].

[100] For instance, Jacques Maritain, relates how at a public hearing to discuss the newly announced Human Rights, a member of the public expressed astonishment that parties representing extremely opposed ideologies had been able to agree on these rights; to which the committee responded: "Yes, we agree about the rights, but on condition that no one asks us why" [As quoted in: 88,89]. Taylor has argued that since then, we have managed to ground these shared human right norms in distinct cultural traditions [90, See also 91].

If 'overlapping consensus' can ground inter-cultural cooperation, could 'incompletely theorized agreements' serve as a foundation for inter-academic cooperation between the near- and long-term AI policy communities? At first sight, there appear to be some caveats. One theoretical objection could argue that an 'incompletely theorized agreement' was not 'meant' to apply to epistemic disagreements, but instead to *normative* disagreements (e.g. relating to communities' sense of justice) and principles. However, the key use of incompletely theorized agreements is not solely that they apply only to normative propositions. Rather, the key use of such agreements is that they allow a given community to bypass (or suspend) any 'intractable' theoretical disagreement that does not appear as if it will be decisively resolved one way or the other in the near term. That can apply to deep philosophical questions as much as to contexts of pervasive scientific uncertainty, especially on questions where it still remains unclear where and how we will procure the information that allows definitive resolution.

Yet this is precisely why an incompletely theorized agreement holds promise. As discussed earlier, debates over AI and the technology's future engage exactly those sources of intractable deep disagreement--on questions of epistemology and norms--that have been the subject of long philosophical debates, unlikely to be resolved any time soon. In addition, the methodology proposed by an incompletely theorized agreement surely does not require that anyone should be expected to dissemble their views in the context of academic and scientific debate. The point is rather that, within the context of urgent societal problems, and given potentially closing time windows of opportunity, it may not just be legitimate, but productive to focus on areas where collaboration is possible to achieve broadly beneficial AI policy outcomes, while debate is ongoing.[101]

[101] To give one example of an incompletely theorized agreement in an adjacent space: Seth Baum has previously argued that policy actions on mitigating 'existential risks' do not require all parties to hold to aggregate utilitarian concerns for the far future; instead, there are at least some policy actions which can help mitigate risks of future catastrophe but which also have co-benefits in the near term, or for many other ethical systems [97].

As such, we want to suggest that these internal epistemic differences, while important to understand the sources of disagreement, may prove less relevant in foreclosing productive areas for cooperation that allow both communities to advance their respective goals in addressing the distinct problems both consider legitimate. While in some cases differential tolerances for uncertainty might undercut philosophical convergence or resolution of these debates, they need not inhibit collaboration on specific policies (such as pursuing strategies for transparency and reliability in AI systems; or in setting ethical AI norms) when these are useful in addressing both near- and long-term concerns.

In summary, we propose an incompletely theorized agreement as a method or principle by which at least some of the underlying theoretical (epistemic, normative or pragmatic) disagreements between the communities working on AI governance can be bracketed, in ways where this impedes neither further scientific debates on these topics (in parallel), but also does not does not halt the development and implementation of productive and much-needed policy cooperation on AI.

5. Conclusion

AI is raising multiple societal and ethical concerns pointing towards a need for impactful and suitable policy measures. At the same time, there is a felt fragmentation between clusters of scholars focusing on 'near-term' AI risks, and those focusing on 'long-term' risks. This paper seeks to map the practical space for inter-community collaboration with a view towards the development of AI policy. We outlined several potential sources of divergence (*epistemic, normative, and pragmatic*). We further presented the novel idea that the constitutional law principle of an 'incompletely theorized agreement' could be used to set aside or suspend these and other disagreements for the purpose of achieving higher order AI policy goals of both communities in selected areas.

This paper does not suggest that communities should fully 'merge' or ignore differences whatever their source may be. To be sure, some policy projects will be relevant to one group within the community but not the other. Indeed, community heterogeneity and diversity is generally a good thing for a scientific paradigm. Instead, the paper proposes to question some possible reasons for conflicting dynamics that could stall positive progress for policy-making, and suggests an avenue for higher order resolution. Most of all, the paper hopes to pragmatically encourage the exploration of opportunities for shared work--on technical research; on shared norm-setting; on forecasting capabilities--and suggested that work on such opportunities, where it is found, can be well grounded through an 'incompletely theorized agreement'.

Bibliography

1. Trajtenberg M. AI as the next GPT: a Political-Economy Perspective. National Bureau of Economic Research; 2018. doi:10.3386/w24245

2. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O. Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. J Artif Intell Res. 2018;62: 729–754. doi:10.1613/jair.1.11222

3. Gruetzemacher R, Whittlestone J. Defining and Unpacking Transformative AI. arXiv:1912 00747 [cs]. 2019. Available: http://arxiv.org/abs/1912.00747

4. Goertzel B, Pennachin C, editors. Artificial General Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. doi:10.1007/978-3-540-68677-4_5

5. Baum SD. Reconciliation between factions focused on near-term and long-term artificial intelligence. AI & Society. 2018;33: 565–572. doi:10.1007/s00146-017-0734-3

6. Baum SD. Medium-Term Artificial Intelligence and Society. Information. 2020;11: 290. doi:10.3390/info11060290

7. Prunkl C, Whittlestone J. Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York NY USA: ACM; 2020. pp. 138–143. doi:10.1145/3375627.3375803

8. Cave S, ÓhÉigeartaigh SS. Bridging near- and long-term concerns about AI. Nature Machine Intelligence. 2019;1: 5–6. doi:10.1038/s42256-018-0003-2

9. Critch A, Krueger D. AI Research Considerations for Human Existential Safety (ARCHES). 2020. Available: http://acritch.com/arches/

10. Drexler KE. Reframing Superintelligence: Comprehensive AI Services as General Intelligence. Oxford: Future of Humanity Institute, University of Oxford; 2019 Jan p. 210. Report No.: 2019-1. Available: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf

11. Christiano P. Prosaic AI alignment. In: AI Alignment [Internet]. 19 Nov 2016 [cited 2 Sep 2020]. Available: https://ai-alignment.com/prosaic-ai-control-b959644d79c2

12. Clark J. Import AI #83: Cloning voices with a few audio samples, why malicious actors might mess with AI, and the industry-academia compute gap. In: Import AI [Internet]. 26 Feb 2018 [cited 23 Jul 2018]. Available: https://jack-clark.net/2018/02/26/import-ai-83-cloning-voices-with-a-few-audio-samples-why-malicious-actors-might-mess-with-ai-and-the-industryacademia-compute-gap/

13. Floridi L, Cowls J, King TC, Taddeo M. How to Design AI for Social Good: Seven Essential Factors. Sci Eng Ethics. 2020. doi:10.1007/s11948-020-00213-5

14. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling Climate Change with Machine Learning. arXiv:1906 05433 [cs, stat]. 2019. Available: http://arxiv.org/abs/1906.05433

15. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. arXiv:1905 00501 [cs]. 2019. Available: http://arxiv.org/abs/1905.00501

16. Calo R. Artificial Intelligence Policy: A Primer and Roadmap. 2017;51: 37. Available: https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Calo.pdf

17. Dafoe A. AI Governance: A Research Agenda. 2018; 52. Available: https://www.fhi.ox.ac.uk/govaiagenda/

18. Müller VC. Ethics of Artificial Intelligence and Robotics. In: Zalta EN, editor. Stanford Encyclopedia of Philosophy. Palo Alto: CSLI, Stanford University; 2020. Available: https://plato.stanford.edu/entries/ethics-ai/

19. Barocas S, Selbst AD. Big Data's Disparate Impact. Calif Law Rev. 2016;671. Available: https://papers.ssrn.com/abstract=2477899

20. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research. 2018. p. 15. Available: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

21. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. arXiv:171000794 [cs]. 2017 [cited 9 Oct 2017]. Available: http://arxiv.org/abs/1710.00794

22. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining

Explanations: An Overview of Interpretability of Machine Learning. arXiv:1806 00069 [cs, stat]. 2019. Available: http://arxiv.org/abs/1806.00069

23. Anderson JM, Kalra N, Stanley K, Sorensen P, Samaras C, Oluwatola TA. Autonomous Vehicle Technology: a Guide for Policymakers. RAND Corporation; 2016. Available: https://www.rand.org/pubs/research_reports/RR443-2.html

24. Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:180207228 [cs]. 2018 [cited 21 Feb 2018]. Available: http://arxiv.org/abs/1802.07228

25. King TC, Aggarwal N, Taddeo M, Floridi L. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Sci. Eng. Ethics. 2018. doi:10.1007/s11948-018-00081-0

26. Hayward KJ, Maas MM. Artificial Intelligence and crime: a primer for criminologists. Crime Media Culture. 2020. doi:10.1177/1741659020917434

27. Stanley J. The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy. American Civil Liberties Union; 2019. Available: https://www.aclu.org/sites/default/files/field_document/061119-robot_surveillance.pdf

28. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Change. 2017;114: 254–280. doi:10.1016/j.techfore.2016.08.019

29. Danaher J. Automation and Utopia: Human Flourishing in a World without Work. Harvard University Press; 2019. Available: https://www.amazon.com/Automation-Utopia-Human-Flourishing-without/dp/0674984242

30. Helbing D, Frey BS, Gigerenzer G, Hafen E, Hagner M, Hofstetter Y, et al. Will Democracy Survive Big Data and Artificial Intelligence? Scientific American. 2017. Available: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/. Accessed 29 May 2017.

31. Nemitz P. Constitutional democracy and technology in the age of artificial intelligence. Philos Trans A Math Phys Eng Sci. 2018;376: 20180089. doi:10.1098/rsta.2018.0089

32. Chesney R, Citron DK. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. Calif Law Rev. 2019;107. Available: https://papers.ssrn.com/abstract=3213954

33. Raso F, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial Intelligence & Human Rights: Opportunities & Risks. Berkman Klein Center for Internet & Society at Harvard University; 2018 Sep. Available: https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf

34. Molnar P. Technology on the margins: AI and global migration management from a human rights perspective. Cambridge International Law Journal. 2019;8: 305–330. doi:10.4337/cilj.2019.02.07

35. Feldstein S. The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression. Journal of Democracy. 2019;30: 40–52. doi:10.1353/jod.2019.0003

36. Danzig R. An irresistible force meets a moveable object: The technology Tsunami and the Liberal World Order. Lawfare Research Paper Series. 2017;5. Available: https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf

37. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford University Press; 2014. Available: https://books.google.nl/books?id=7_H8AwAAQBAJ

38. Russell S. Human Compatible: Artificial Intelligence and the Problem of Control. Viking; 2019. Available: https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem-ebook/dp/B07N5J5FTS

39. Parson E, Re R, Solow-Niederman A, Zeide A. Artificial Intelligence in Strategic Context: An Introduction. PULSE, UCLA School of Law; 2019. Available: https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/

40. Kaplan S, Radin J. Bounding an emerging technology: Para-scientific media and the Drexler-Smalley debate about nanotechnology. Soc Stud Sci. 2011;41: 457–485. Available: https://www.jstor.org/stable/41301944

41. Taddeo M, Floridi L. How AI can be a force for good. Science. 2018;361: 751–752. doi:10.1126/science.aat5991

42. Etzioni O. How to know if artificial intelligence is about to destroy civilization. MITS Technol Rev. 2020. Available: https://www.technologyreview.com/2020/02/25/906083/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/

43. Metz C. The AI Threat Isn't Skynet. It's the End of the Middle Class. WIRED. 2 Oct 2017. Available: https://www.wired.com/2017/02/ai-threat-isnt-skynet-end-middle-class/. Accessed 15 Feb 2017.

44. Geist EM. Is artificial intelligence really an existential threat to humanity? - Bulletin of the Atomic Scientists. In: Bulletin of the Atomic Scientists [Internet]. 9 Aug 2015 [cited 4 Dec 2018]. Available: https://thebulletin.org/2015/08/is-artificial-intelligence-really-an-existential-threat-to-humanity/

45. Russell S, Dafoe A. Yes, the experts are worried about the existential risk of artificial intelligence. In: MIT Technology Review [Internet]. 2 Nov 2016 [cited 26 Feb 2017]. Available: https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/

46. Baum SD. Countering Superintelligence Misinformation. Information. 2018;9: 244. doi:10.3390/info9100244

47. Future of Life Institute. AI Safety Myths. In: Future of Life Institute [Internet]. 2016 [cited 26 Oct 2017]. Available: https://futureoflife.org/background/aimyths/

48. Avin S, Amadae SM. Autonomy and machine learning at the interface of nuclear weapons, computers and people. In: Boulanin V, editor. The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Stockholm International Peace Research Institute; 2019. doi:10.17863/CAM.44758

49. Grace K. Likelihood of discontinuous progress around the development of AGI. In: AI Impacts [Internet]. 23 Feb 2018 [cited 26 Mar 2018]. Available: https://aiimpacts.org/likelihood-of-discontinuous-progress-around-the-development-of-agi/

50. Adamczewski T. A shift in arguments for AI risk. Fragile Credences. 2019. Available: https://fragile-credences.github.io/prioritising-ai/

51. Müller VC, Bostrom N. Future Progress in Artificial Intelligence: A Survey of

Expert Opinion. In: Müller Vincent C., editor. Fundamental Issues of Artificial Intelligence. Berlin: Synthese Library; 2016. Available: http://www.nickbostrom.com/papers/survey.pdf

52. Baum SD, Barrett AM, Yampolskiy RV. Modeling and Interpreting Expert Disagreement About Artificial Superintelligence. Informatica. 2017;41. Available: http://www.informatica.si/index.php/informatica/article/view/1812

53. Armstrong S, Sotala K. How We're Predicting AI--Or Failing To. In: Romportl J, Ircing P, Zackova E, Polak M, Schuster R, editors. Beyond AI: Artificial Dreams. Pilsen: University of West Bohemia; 2012. pp. 52–75. Available: https://intelligence.org/files/PredictingAI.pdf

54. Armstrong S, Sotala K, OhEigeartaigh SS. The errors, insights and lessons of famous AI predictions – and what they mean for the future. J Exp Theor Artif Intell. 2014;26. Available: http://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf

55. Ord T, Hillerbrand R, Sandberg A. Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. arXiv:08105515 [physics]. 2008;13.0: 191–205. doi:10.1080/13669870903126267

56. Bostrom N. Existential Risk Prevention as Global Priority. Global Policy. 02/2013;4: 15–31. doi:10.1111/1758-5899.12002

57. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety. arXiv:160606565 [cs]. 2016 [cited 13 May 2017]. Available: http://arxiv.org/abs/1606.06565

58. Krakovna V, Uesato J, Mikulik V, Rahtz M, Everitt T, Kumar R, et al. Specification gaming: the flip side of AI ingenuity. Deepmind. 2020. Available: https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity

59. Kumar RSS, Brien DO, Albert K, Viljöen S, Snover J. Failure Modes in Machine Learning Systems. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1911.11034

60. Turner AM. Optimal Farsighted Agents Tend to Seek Power. arXiv [cs.AI]. 2019. Available: http://arxiv.org/abs/1912.01683

61. Crawford K, Calo R. There is a blind spot in AI research. Nature News. 2016;538: 311. doi:10.1038/538311a

62. Bostrom N, Dafoe A, Flynn C. Public Policy and Superintelligent AI: A Vector Field Approach. In: Liao SM, editor. Ethics of Artificial Intelligence. Oxford University Press; 2020. Available: http://www.nickbostrom.com/papers/aipolicy.pdf

63. Brooks R. The Seven Deadly Sins of Predicting the Future of AI. 7 Sep 2017 [cited 13 Sep 2017]. Available: http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

64. Sutton R. The Bitter Lesson. 2019. Available: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

65. Brooks R. A Better Lesson. 2019. Available: https://rodneybrooks.com/a-better-lesson/

66. Marcus G. Deep Learning: A Critical Appraisal. arXiv:1801 00631 [cs, stat]. 2018. Available: http://arxiv.org/abs/1801.00631

67. Hernandez-Orallo J, Martınez-Plumed F, Avin S, Whittlestone J. AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues. Santiago de Compostela, Spain; 2020. p. 8. Available: http://ecai2020.eu/papers/1364_paper.pdf

68. Manheim D, Garrabrant S. Categorizing Variants of Goodhart's Law. arXiv:1803. 04585 [cs, q-fin, stat]. 2018. Available: http://arxiv.org/abs/1803.04585

69. Thomas R, Uminsky D. The Problem with Metrics is a Fundamental Problem for AI. arXiv [cs.CY]. 2020. Available: http://arxiv.org/abs/2002.08512

70. McDonald H. Home Office to scrap "racist algorithm" for UK visa applicants. The Guardian. 4 Aug 2020. Available: http://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants. Accessed 2 Sep 2020.

71. Illinois General Assembly - Full Text of HB2557. 2019 [cited 2 Sep 2020]. Available: https://www.ilga.gov/legislation/fulltext.asp?DocName=&SessionId=108&GA=101&DocTypeId=HB&DocNum=2557&GAID=15&LegID=&SpecSess=&Session=

72. SB-1001 Bots: disclosure. In: California Legislative Information [Internet]. 2018 [cited 2 Sep 2020]. Available:

https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

73. Belfield H. Activism by the AI Community - Analysing Recent Achievements and Future Prospects. Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics and Society 2020. 2020.

74. Perry B, Uuk R. AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk. Big Data and Cognitive Computing. 2019;3: 26. doi:10.3390/bdcc3020026

75. Hallsworth M, Parker S, Rutter J. Policymaking in the real world: evidence and analysis. Institute for Government; 2011 Apr. Available: https://www.instituteforgovernment.org.uk/sites/default/files/publications/Policy%20making%20in%20the%20real%20world.pdf

76. Raji ID, Buolamwini J. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. 2019. p. 7. Available: http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf

77. McNamara A, Smith J, Murphy-Hill E. Does ACM's code of ethics change ethical decision making in software development? Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018. Lake Buena Vista, FL, USA: ACM Press; 2018. pp. 729–733. doi:10.1145/3236024.3264833

78. Cleek MA, Leonard SL. Can corporate codes of ethics influence behavior? J Bus Ethics. 1998;17: 619–630. doi:10.1023/A:1017969921581

79. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: Association for Computing Machinery; 2020. pp. 33–44. doi:10.1145/3351095.3372873

80. Baum SD. On the promotion of safe and socially beneficial artificial intelligence. AI Soc. 2016 [cited 13 May 2017]. doi:10.1007/s00146-016-0677-0

81. Shevlane T, Dafoe A. The Offense-Defense Balance of Scientific Knowledge:

Does Publishing AI Research Reduce Misuse? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery; 2020. pp. 173–179. doi:10.1145/3375627.3375815

82. Whittlestone J, Ovadya A. The tension between openness and prudence in AI research. arXiv:1910 01170 [cs]. 2020. Available: http://arxiv.org/abs/1910.01170

83. Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J, et al. Release Strategies and the Social Impacts of Language Models. arXiv:1908 09203 [cs]. 2019. Available: http://arxiv.org/abs/1908.09203

84. Bostrom N. Strategic Implications of Openness in AI Development. Glob Policy. 2017 [cited 18 Feb 2017]. doi:10.1111/1758-5899.12403

85. Cave S, Coughlan K, Dihal K. "Scary Robots": Examining Public Responses to AI. Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics and Society 2019. 2019. p. 8. Available: http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_200.pdf

86. Sunstein CR. Incompletely Theorized Agreements. Harv Law Rev. 1995;108: 1733–1772. doi:10.2307/1341816

87. Sunstein CR. Incompletely Theorized Agreements in Constitutional Law. Soc Res . 2007;74: 1–24. Available: http://www.jstor.org/stable/40971887

88. Sunstein CR. Holberg Prize 2018, Acceptance Speech. Holberg Prize 2018; 2018 Jun 7; Bergen, Norway. Available: https://www.holbergprisen.no/en/cass-sunsteins-acceptance-speech

89. UNESCO. Human Rights: Comments and Interpretations. 1948. Available: https://unesdoc.unesco.org/ark:/48223/pf0000155042

90. Taylor C. Conditions of an Unforced Consensus on Human Rights. Bangkok; 1996. Available: https://www.iilj.org/wp-content/uploads/2016/08/Taylor-Conditions-of-an-Unforced-Consensus-on-Human-Rights-1996.pdf

91. ÓhÉigeartaigh SS, Whittlestone J, Liu Y, Zeng Y, Liu Z. Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. Philos Technol. 2020. doi:10.1007/s13347-020-00402-x

92. Ruger JP. Pluralism, Incompletely Theorized Agreements, and Public Policy. Health and Social Justice. Oxford: Oxford University Press; 2009. doi:10.1093/acprof:oso/9780199559978.003.0004

93. Rawls J. Political Liberalism. Columbia University Press; 1993.

94. Benjamin M. The Value of Consensus. Society's Choices: Social and Ethical Decision Making in Biomedicine. National Academy Press; 1995. doi:10.17226/4771

95. Søraker JH. The role of pragmatic arguments in computer ethics. Ethics Inf Technol. 2006;8: 121–130. doi:10.1007/s10676-006-9119-x

96. Hongladarom S. Intercultural Information Ethics: A Pragmatic Consideration. In: Kelly M, Bielby J, editors. Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro. Wiesbaden: Springer Fachmedien; 2016. pp. 191–206. doi:10.1007/978-3-658-14681-8_11

97. Baum SD. The Far Future Argument for Confronting Catastrophic Threats to Humanity: Practical Significance and Alternatives. Futures. 2015;72: 86–96. Available: http://sethbaum.com/ac/2015_FarFuture.html

Chapter V

Foundations for the future: Institution building for the purpose of artificial intelligence governance

……..

Stix, C. Foundations for the future: institution building for the purpose of artificial intelligence governance. *AI Ethics* 2, 463–476 (2022). https://doi.org/10.1007/s43681-021-00093-w.

……..

Introduction

The increasing capabilities of artificial intelligence (AI) systems,[102] and their manifold applications throughout society, have given rise to a range of governance concerns. New policies are needed to ensure the safe and reliable use and robust behavior of these systems (Müller, 2020; Calo, 2017; Ulnicane et al., 2020), in accordance with fundamental and human rights (Raso et al., 2018) and ethical principles (Schiff et al., 2020). Recently, governments around the world have begun to approach the governance of AI through multiple levers (Cyman, Gromova, and Juchnevicius, 2021). One timely example is the EU's horizontal regulation published in April 2021 which puts forward a regulatory framework for high-risk AI systems that are brought into or put on the Single Market (European Commission 2021). Another timely example is the Trade and Tech Council[103] established in the

[102] This paper follows the definition of AI put forward by the European Commission's High Level Expert Group on AI (AI HLEG, 2019).
[103] See: https://ec.europa.eu/commission/presscorner/detail/en/IP_21_2990.

summer of 2021 between the United States and the European Union. One of the key policy areas for this Tech and Trade Council is to cooperate on the development of suitable standards for AI. Policy levers such as regulation or standardization will play a crucial role when it comes to shaping the design, development and technical benchmarking of future AI systems on an international scale (Cihon, 2019; Bradford, 2020). The field of AI governance is relatively nascent and, as a result, there exist few dedicated specialist governmental institutions exclusively dedicated to supporting many of these initiatives. In order to properly develop, support and implement new AI governance efforts, such as legislative frameworks for AI, it is likely that a number of new institutions will need to be established over the coming years (Turner, 2018; Almeida et al., 2021). It is therefore worthwhile to step back and think cautiously about the junctures we are about to lay (Pierson, 2000) for the field of AI governance when developing these new institutions.

There are broadly two types of institutions that one could investigate: those that exist and may be adapted (i.e. either they organically evolve over time, *or* are 'repurposed' see (Alter & Raustiala, 2018; Kunz & ÓhÉigeartaigh, 2019; Stahl et al., 2021), and those that do not exist yet (but will eventually come into existence to fill the void that new governance measures have created and support them). This paper focuses on the latter and with a particular eye to those institutions set up by governments. This does not mean to make a judgment as to the relative importance of the work of NGOs in shaping AI governance, rather it serves to narrow the focus of this paper's investigation.

In doing so, it picks up from recent academic calls for an international governance coordinating committee for AI (Wallach & Marchant, 2018), for an international regulatory agency for AI (Erdélyi & Goldsmith, 2018) and for a G20 coordinating committee for AI governance (Jelinek, Wallach & Kerimi, 2020). Moreover, it takes up suggestions by scholars that at times it might be easier to add new institutions than to change or dismantle existing ones, if the latter cannot handle problems

adequately (Alter & Raustiala, 2018). Building on these calls and drawing on existing scholarship and insights, it addresses itself to those individuals who will be involved in setting up new institutions and those who are interested in conducting further research on pragmatic institution building for AI governance. The ambition of this paper is to empower them to make timely and suitable interventions in upcoming institution building efforts.

First, the paper outlines the rationale for deepening the investigation of institution-building for AI governance now, as well as the limitations to this paper's investigation. Subsequently, it puts forward three axes that contribute to the make-up of an institution: (i) *purpose*: what an institution is meant to do; (ii) *geography*: who are the members and what is the scope of jurisdiction; and, (iii) *capacity*: what and who makes up the institution. It breaks the latter down into infrastructural and human elements. Furthermore, under each subsection in 2.1, it highlights the pros and cons of various institutional roles and briefly frames what they could look like in practice. It does so by placing these roles in a European context and sketches a potential European AI Agency with them. Finally, conclusions and future research directions are put forward. Overall, the paper highlights pragmatic governance considerations and refrains from making normative assessments of these various institutional setups.

1. Motivation, urgency and limitations

Novel AI specific governance institutions working on soft governance mechanisms, that is non-binding rules, have already come into existence. For example, several intergovernmental fora such as the G7 or the OECD engage on mechanisms such as shared ethical principles (OECD AI Principles). One of the European Commission's foreign policy mechanisms, the International Alliance for a Human Centric Approach to AI[104] which cooperates with like-minded countries such as Singapore,

[104] See:
https://digital-strategy.ec.europa.eu/en/funding/international-alliance-human-centric-approach-artific

Japan or Canada to build on the European vision of a 'trustworthy AI' (AI HLEG, 2019). The Global Partnership on AI, through its 15 founding members and various dedicated working groups, aims to strengthen shared norms for AI through coordinated efforts. The OECD's ONE-AI working groups, adjacent to the OECD AI Policy observatory, supports the OECD recommendations to its members, ranging from policy measures to implementing trustworthy AI and compute. Prima facie these new fora appear well placed to support and engage in soft governance efforts.

At the same time there is mounting pressure to develop and implement stronger and more binding AI governance mechanisms than those covered by ethical principles, shared norms, and multi-stakeholder proposals. This pressure comes in a number of forms, such as from a societal viewpoint, accounting for the range of issue areas AI can cause (Butcher & Beridze, 2019; Whittlestone, Arulkumaran, and Crosby 2021) and a technological viewpoint, accounting for the increasing capacity of AI and the speed of its development (Grace et al., 2018). In response, for example at the trans-national level, the Council of Europe is currently examining what a legal framework for AI (international law) built on the Council of Europe's values could look like through its CAHAI committee (Ad Hoc Committee on Artificial Intelligence).[105] Overall, as countries move towards harder governance efforts, such as regulation (European Commission, 2021), standardization[106] or certification they are likely to require increasingly specialized institutions.

Moreover, as AI governance efforts soar and more coordination, action and policy proposals become necessary within a nation as well as at an international level, it is likely that there will be a need for more (and more specialized) governmental

ial-intelligence#:~:text=International%20alliance%20for%20a%20human%2Dcentric%20approach%20to%20Artificial%20Intelligence,-Opening%3A%2006%20November&text=human%2Dcentric%20AI., common%20principles%20and%20operational%20conclusions.

[105] See: https://www.coe.int/en/web/artificial-intelligence/cahai.

[106] See for example work on AI standards by the ISO/IEC JTC 1/SC 42 Committee. Available here: https://www.iso.org/committee/6794475.html.

agencies to handle an increasingly diverse portfolio on top of existing files. Especially in cases where specialization on a particular AI governance angle or AI application in a particular sector is required, and if that area is set to increase in volume, it is probable that the setting up of a specialized institution might be beneficial. It might be overall quicker, cheaper and more effective to build a new institution from scratch that is 'fit for purpose' rather than exert time, effort and political goodwill to change the structure of an existing institution.

Far from being a hypothetical proposition, this matters now: the EU is laying out extensive plans to set up new institutions for AI governance in the coming years (AI Act, 2021), NATO is putting forward plans for a civil-military Defense Innovation Accelerator for the North Atlantic, and many others will follow suit.

In April 2021, the European Commission put forward their horizontal risk-based regulation for AI, the AI Act (European Commission, 2021). In that AI Act, they described what form a regulatory framework for AI would look like in the EU and indicated a number of institutions that would partake in the fulfillment process of the regulatory requirements. While some of these already exist and will likely have their remit extended, such as market surveillance authorities, others, such as new institutions responsible for testing, certifying and inspecting AI systems (notified conformity bodies) are likely to be established to cope with an increasing demand once the AI Act is enforced. Moreover, while the AI Act does not foresee one single supervisory Agency for AI in the EU, other EU institutions like the European Parliament have advocated for such a new Agency. The 2017 resolution on 'Civil Law Rules on Robotics' (European Parliament, 2017) and the 2019 report on 'A comprehensive European industrial policy on artificial intelligence and robotics' (European Parliament, 2019) both called for the establishment of a European AI Agency that would be in charge of supplying technical, legal and ethical expertise and intervention. Given the sheer scope the new horizontal legislation for AI will introduce and the number of associated institutions, networks and processes that

need to be either set up or managed, and given this political direction from the European Parliament, it is not unlikely that a European AI Agency will be on the horizon within the coming 10 years. This institution is likely enough to be established, yet far enough in the future for there to be meaningful interventions to its structure. It will be used as a practical example under each axis in Section 3.1. to demonstrate how various considerations could play out.

Different institutional set-ups will yield different path dependencies. Those that do get put into motion over the coming years are likely to be especially critical (Stix & Maas, 2021) because they form the lens through which governments will be able to interact with progressing AI technologies and enact appropriate AI governance measures. This wouldn't merit outsized concern, if it were possible to clearly forecast AI progress across various sectors and if there was common expert consensus about the development path of AI over the coming decades. However, even experts struggle to agree (Grace et al., 2018; Müller & Bostrom, 2016) and given the ubiquitous nature of AI it is difficult, if not impossible, to forecast now what AI governance frameworks might be needed in the near or far future. In addition, existing institutions, their policy texts and legal frameworks are historically often drawn from in moments of unexpected change (Stix and Maas 2021), because it can be quicker to react with what exists than develop something new.

Once new institutions are established they are difficult to change (Sanders, 2008). Equally, it is difficult to adapt larger international institutions to changing issues of concern within the landscape (Morin et al., 2019) or to shift from the processes that were originally conceived to different mechanisms and fluctuating missions (Baccaro & Mele, 2012). This all goes to say that, where new institutions may be the vehicle of choice to action current policy measures, in the future, the composition of these new institutions will itself precede and, crucially, *constrain* or *catalyze* future governance approaches. Early stage decisions to establish new institutions, or the

choice to forego such new institutions (Cihon, Maas, & Kemp, 2020), are all likely to have a downstream, or lock-in, effect on the efficiency of government measures and on the field as a whole. It is therefore timely and important to think about institution building as one of the critical path dependencies we are developing now.

Prior to delving into a deeper investigation of institutional set-ups and associated considerations, we place the work within the existing literature and outline the scope of the work set out in the remainder of the article. The paper builds on and contributes to existing literature on the topic of institution building (Goodin, 1998; Koremenos, Lipson, and Snidal, 2001a; Koremenos, Lipson, and Snidal, 2001b) by focusing on the topic of institution building and design for AI governance interventions specifically. In particular, the paper adopts the lens of historical institutionalism (Thelen, 1999), suggesting that for AI governance new institutions will be built dependent on the current promise and peril of (and the associated mechanisms to enable or minimize) AI. Newly established AI governance institutions will themselves shape political interactions as those interactions will be a result of the newly established institutions and their processes in turn (Sanders, 2008). The question of how to set up AI governance institutions therefore merits further scholarly investigation.

Accordingly, Section 2 puts forward a selection of axes that need to be considered in building new AI governance institutions: (i) *purpose*, (ii) *geography*, and (iii) *capacity*. These provide a framework for envisioning different types of institutional designs and their relative merits and shortcomings.

This preliminary set of axes does not claim to be exhaustive. Neither can each individual investigation fully account for the range of interactions, interdependencies and considerations that may (or should) exist. Inevitably, the creation of institutions for AI governance will have a political and geopolitical dimension. Angles such as those of cooperation theory are relevant to that

dimension, but fall outside of the scope of this paper.

2. The main axes: purpose, geography and capacity

The following investigation maps the axes of *purpose*, *geography* and *capacity*, with particular focus paid to *purpose*. It serves to outline a first framework towards institution building for AI governance. How the aspects are combined, what is favored and what isn't, will depend on the values, difficulties and models of the future/AI governance worldview each reader holds, as well as on the political, societal and economic climate those setting out to establish a new institution will find themselves in.

2.1. Purpose: What is it meant to do?

Among the first things that will need to be decided upon is the *purpose* of the new institution. That is: *what is it meant to do*? Prima facie, this might seem like a straightforward question, yet the following paragraphs will evidence why neither the question nor the answer(s) are. Each subsection introduces the outline of a role an institution for AI governance could take: *coordinator*, *analyzer*, *developer* and *investigator*. Each subsection introduces the relevant role, followed by examples of existing institutions that match this model (where appropriate), a discussion of some relevant considerations, and concludes with transferring the explored role to the case of a potential European AI Agency, demonstrating what shape this might take in practice.

Finally, this section concludes with a short account of hybrid cases between the various roles, briefly accounting for how they might intersect or support each other in practice.

2.1.a. The coordinator institution

Current proposals by academic researchers (Wallach & Marchant, 2018; Jelinek, Wallach, & Kerimi, 2020) and governments (such as the European AI Board, European Commission, 2021) often suggest something akin to what this paper terms a *coordinator*: an institution whose purpose it is to coordinate between a number of actions, policy efforts or norms.

What does it do? Coordinator institutions could, for example, work with the rising number of ethical guidelines (Zeng, Lu, & Huangfu, 2018) and attempt to operationalize them more clearly. They could also serve as an umbrella organization to coordinate activities across like-minded nations (Erdélyi & Goldsmith, 2018), helping different groups to learn lessons from one another and avoid duplicating efforts.

Moreover, there is a rising number of relevant but often uncoordinated efforts to tackle certain aspects relevant to AI governance, such as certification schemes (Winter et al., 2021), testing procedures (Brundage et al., 2020), or setting out shared definitions of 'meaningful human control' across diverse contexts. These too could benefit from coordination to increase efficiency, coherence and policy impact. A coordinator institution (others have referred to similar roles as 'orchestrator' cf. (Abbott & Snidal, 2010; Abbott & Genschel, 2000), would focus on encouraging the exchange and synchronization between institutions or efforts, amplifying and streamlining work done elsewhere. It could fill a current gap 'in the market' of some aspects of AI governance, stemming issues that arise from quickly developing and independent streams of efforts, and combat a fragmented landscape as well as norm conflict (Garcia, 2020).

What models of coordinator organizations exist? Intergovernmental organizations are the closest example of a coordinator role for the purpose of AI governance. While they are much more complex and encompass some roles discussed later in this

section, intergovernmental organizations such as the United Nations and organizations covered thereunder such as the ITU; or the G20 or the NATO can be seen as coordinators. They predominantly act to coordinate efforts, strategies and common interests under international law and are composed of sovereign states (actors) that share mutual interests or seek a neutral platform for discussion and exchange. From a European perspective, the European Data Protection Board coordinating between various national data protection agencies could be seen as a coordinator. *What are relevant considerations?*

Power. One of the benefits of a coordinator could be that it serves as a relatively neutral environment within which various members (e.g. groups, nations) are able to discuss and develop broad agreement on shared initiatives. It may also serve to alleviate the workload of its members by acting as a 'quasi back office' supporting synchronization, distribution and organization of relevant work streams.

The coordinator's main purpose is to serve to coordinate for its members and not to hold an independent political agenda. Nevertheless, it should not be mistaken as being completely powerless or unable to (inadvertently) shape the prominence of certain governance efforts.

Depending on how it is set up a coordinator is either *independent* to its members, or the coordinator *is* the members. Similarly, depending on the procedures established through which it can act, either version would have decision making power over the approval of new members and the decision making power over new coordination efforts within its remit. In either case, whether willingly or unwillingly, the coordinator will inherently only amplify the mutual interest of its members vis-a-vis the international political stage.

This can lead to exclusion and oversight, uninformed or ill-informed choices to develop or deploy AI systems which may negatively impact nations or groups that

do not form part of the coordinator (Hagerty & Rubinov, 2019), or an amplification of approaches that might be suited for some, but unsuited for the entirety of the international community (Schiff et al., 2020).

Moreover, the act of coordination itself comes with trade-offs: navigating a common line among efforts or approaches might result in some falling outside of the plotted line. This would translate into a decrease of support from the coordinator for those efforts given they do not match the majority line. On the other side of the coin, those efforts or approaches with more resources to start with are likely to have a higher ability of dominating the common line, or even to hedge their bets and participate in parallel in more than one coordination regime.

Access. Questions regarding membership procedure should be addressed while setting up the institution or be clearly related to the topic of activities the coordinator is established for. For example, when the EU discusses coordination and cooperation with 'like-minded countries' through its external policy vehicle, the International Alliance for Human Centric AI, then this narrows the scope of engagement to those countries that follow a similar adherence to fundamental and human rights, and values. While access to the coordinator should be possible to all relevant actors in order to ensure diversity and representation in value and opinion, quasi open access does come with a trade off. For a coordinator, more members might result in more time spent on early stage coordination interventions and mutual understandings, before time can be spent on actioning those understandings and streamlining efforts.

The coordinators' ability to fulfill their ultimate duty and to add value as an institution becomes increasingly difficult the thinner the coordinator is spread across increasing efforts, novel angels and shifting needs. This could also contribute to slower reactions to and adaptations to potentially quickly shifting technical landscapes (Marchant, Allenby, & Herkert, 2011).

Focus. The focus of the coordinator, and its actions should be timely and appropriate. The focus will likely be dictated by the initial group of members. Timely and appropriate focus also encompassess a certain degree of agility. This might be especially important to bear in mind if and when the coordinator expands, and in light of foreseen and unforeseen shifts and needs within the AI governance landscape. For example, in response to an unexpected crisis such as the COVID-19 pandemic previously known issues such as those of privacy and security in light of technological tools might become imminently pressing and require fast coordinated responses (Tzachor et al., 2020) that cross national boundaries.

Depending on the flexibility the coordinator is endowed with, either independently from its members or through e.g. voting procedures from its members, the initial set up could have an outsized impact on the future agility and responsiveness of the coordinator. It could provide an 'first mover advantage' to those that set up the institution and set its first directions.

How could this look like in practice? A future AI Agency in the EU might take the role of a coordinator. In practice, this could mean that the 27 EU member states would be the founding members of that agency, as they are the only actors that are directly involved with the EU's governance efforts such as regulation. In that sense, it would be sufficiently inclusionary and is unlikely to need to structure elaborate access procedures in the near future (until a new country joins the EU). While its scope may not be stretched through access of new members, it might nevertheless stretch itself thin quickly on actions, unless there is a narrowly defined scope for the coordinator EU AI Agency. The large number of members might also lengthen reaction times under unforeseen circumstances.[107]

[107] See Commission President von der Leyen's statement describing the EU as a tanker, whereas an independent country – such as the UK – can act as a speedboat: https://www.theguardian.com/society/2021/feb/05/ursula-von-der-leyen-uk-covid-vaccine-speedboat-eu-tanker.

Currently, a multitude of efforts are coordinated between member states through the European Commission and under the helmet of the Coordinated Plan on AI (European Commission, 2018; European Commission, 2021). In addition, new challenges that a coordinator could tackle will arise out of the AI Act (European Commission, 2021) and its implementation across the EU. The coordinator EU AI Agency could pick up coordination efforts under the Coordinated Plan on AI (European Commission, 2018; European Commission, 2021) such as aligning AI strategies, pooling resources and strengthening the ecosystem between member states. Or, it could act as a coordinator EU AI Agency for the AI Act supporting the coherent implementation, enactment and functioning of the horizontal regulation for AI.

Given aforementioned considerations, the highest benefit might be derived if its focus is clear cut and does not overlap with existing efforts such as those already led by the European Commission. As such, coordination on the AI Act (European Commission, 2021) and either (i) coordination between (a) all relevant institutions within the member states' landscape (e.g. national supervisory authority, notifying authorities or notified bodies), (b) between a subset of similar institutions within the member states' landscape such as notified conformity assessment bodies; or, (ii) on specific topics, could derive the highest benefit.

It would mean that there is one coordinator that is tasked with monitoring, implementing and supervising relevant activities with regards to the AI Act and/or a number of smaller corresponding bodies across the EU. This could strengthen information exchange and increase the speed at which scope specific problems can be identified and addressed, ultimately supporting the higher order goal of a well functioning regulatory framework for AI.

2.1.b. The analyzer institution

Another role for a future AI governance institution could be that of an analyzer.

What does it do? An analyzer could fulfill several roles: it could serve to map existing efforts and identify gaps across various governments. The European Commission, for example, undertook such mapping efforts in the Coordinated Plan on AI: 2021 review (European Commission, 2021) to identify where specific AI-relevant measures had not yet been implemented or needed to be established across the EU. It could compile data sets and information on the technical landscape and sketch technological trajectories. The AI Index,[108] for example, tracks and publishes data corresponding to progress within various AI applications and research areas. It could collate relevant information about the opportunities and risks associated with the use of AI within specific sectors along the lines of the Center for Data Ethics and Innovation's (CDEI) AI Barometer[109] work. In short, an analyzer draws new conclusions from qualitative and quantitative information.

What models of an analyzer exist? In addition to the aforementioned examples, where in the case of the European Commission and the CDEI the analyzer role of those institutions is housed within a larger institutional framework holding multiple roles, a good example of an analyzer institution is the European Parliament's Think Tank.[110] It provides studies, briefing and in-depth analyses on a variety of topic areas to the members of the European Parliament and makes them publicly available. These can be requested from the European Parliament or developed subsequent to a particularly pertinent happening. In doing so, the European Parliament Think Tank supports the well-functioning of the European Parliament, the speed at which relevant decisions can be made and heightens the understanding of participants on relevant key topics. Another example would be the

[108] See: https://hai.stanford.edu/research/ai-index-2021.
[109] See: https://www.gov.uk/government/publications/cdei-ai-barometer.
[110] See: https://www.europarl.europa.eu/thinktank/en/home.html.

European Commission's Joint Research Center's AI Watch[111] which monitors, and provides high-level analyses about, research, industrial capacity and policy initiatives across the EU to inform the European Commission on policy decisions.

What are relevant considerations? The role of an analyzer is more active than that of a coordinator, in that it interferes more directly with the governance or policy making process by way of providing crucial information that can inform and shape those decision making processes.

What or who does it respond to? One key consideration is whether an analyzer is established in response to an identified need within the governance landscape or whether it is established without a clear angle in mind but as an independent institution to provide services to advise decision-makers. In the first case, the analyzer would be directly driven by its scope and gain its direction from the event and those who chose to establish it as a response. The second case is a lot broader and such an institution could, in principle, allow for a more diverse range of ad hoc analyzes, be that on specific sectors such as the automotive sector, the financial sector or the healthcare sector, or on specific topic areas such as compute or data. The European Parliament's Think Tank would fall into this category. Instead of serving to inform one particular direction, it adapts to the shifting needs of those who established it or those who it is supposed to inform, be that high-level individuals, governmental agencies or entire governments. It might therefore be able to cover more breadth. However, one trade-off might be in-depth subject expertise. An analyzer focused on one specific topic area could quickly become an expert institution in that field, whereas an analyzer focused on multiple topics might need to draw on outside expertise for particular areas which could cause additional time and effort.

[111] See: https://knowledge4policy.ec.europa.eu/ai-watch/about_en.

Independence. Using the aforementioned examples, it would appear that an analyzer can either be (i) dependent on those that demand its work, or (i) independent to those that demand its work. In both cases, though to a varying degree, the quality of the work of the analyzer will depend on the range of relevant information it has access to. In the case of working with publicly available information this ought not to pose a problem, however, if it concerns information that is predominantly derived from non-public sources then the analyzer is dependent on the information provider to ensure that it is complete, accurate and comprehensive. Where the information provider may overlap with the group that asks the analyzer to conduct work on their behalf this can become tricky as the result of the analysis might be (inadvertently) shaped by the actor that requests the work. This could even hold true where no information is exchanged but where the framing of the work, or question, is given by the actor requesting the service and not developed by the analyzer independently, (inadvertently) narrowing the scope of research and shaping its direction. Moreover, in these cases, existing narratives might be amplified instead of empowering the discovery of novel, equally pressing ones (Kak 2020).

Shape of the final product. Beyond the process which the analyzer undergoes to develop its final piece of work, another consideration is the actual shape of that work.

That means, the final product could be as 'shallow' and 'inactive' as identifying shared issues and solutions, or more towards a 'deeper' and more 'active' product by providing coherent policies or sets of requirements to be implemented based on the work undertaken. The depth of analysis and the ease with which it can be actioned or implemented matters to the impact the final product of the analyzer will have. Whether these suggestions are binding to the actor(s) that requested the analyzer to undertake the work, and the degree to which they need to act based on the information received will differ. On that spectrum, between non-binding suggestions

with no need to act on them and either binding suggestions or those where action and change are a clear expectation and requirement, the analyzer's importance within the AI governance ecosystem will differ.

Put differently, the higher the likelihood of its work being adopted in decision making processes will be, the more influential it is and therefore this will likely affect previous questions surrounding its (political or financial) independence. In a situation where policymakers and governments rapidly look to develop new approaches towards the governance of AI, analyzers hold some non-negligible power. For one, they can influence those looking towards finding solutions, measures and new information under the pressure of addressing rapid tech development in comparison to slow policy making filling a void in current policy making (Marchant, Allenby, & Herkert, 2011).

How could this look like in practice? For the hypothetical case of a European AI Agency, one version of an analyzer role could be independent from those that request its work. It could serve as a third party compiling data on specific efforts which it can then offer to various institutions within the ecosystem in the form of recommendations in order to change or amend their actions when it identifies pain points. For example, in the case of the implementation of the horizontal regulation for AI (European Commission, 2021) if the analyzer identifies that assessments of some high-risk AI systems are slow across the EU for a specific subset of the third party conformity assessment procedure it can recommend to increase the number of notified conformity assessment bodies on that topic. Depending on how it is set up, its work could be both binding and non-binding which would influence its ability to meaningfully shape the pain points it identifies.

On the other hand, an analyzer role for a European AI Agency could also be dependent on those that have set it up, responding to various different requests for its work from a range of actors, such as surveys on AI uptake, statistics on overall

adherence to the regulatory framework or divergences between member states. In this case, the analyzer would likely respond to the European Commission and member states or a subset of the institutions across the EU that are involved in the regulatory framework, which might narrow its scope to discover novel pain points in accordance.

In each of these two cases, the motivation of the analyzer's work will be different and either depend on the priorities it has identified for itself or on the priorities of the groups it represents.

Finally, a European AI Agency could also take the role of an AI observatory. One version could be an AI observatory that is set up by EU institutions and the governments of the member states but sufficiently independent to provide meaningful insight into the actual state of AI development and deployment across the EU in addition to existing work. Such an AI observatory could also double down on quantitative analysis only, to gain a competitive advantage over existing work in that space in the EU. It could track, measure and eventually forecast the progression of various AI technologies, map relevant aspects of their supply chains or components to develop AI systems (e.g. compute), and monitor capacities and impacts of deployed AI systems across the EU.

2.1.c. The developer institution

Most versions of an analyzer institution will stop short of coming up with clear proposals to implement and action its analysis.

What does it do? Where an analyzer might identify gaps in the technical landscape, a developer would structure policy proposals to close the gaps. Where an analyzer might map adherence to certain principles or ambitions, a developer would argue

which ones are more important and give advice on how to strengthen those. Where an analyzer would suggest that there is a lack of sufficient institutions to account for a certain procedure needed within the ecosystem, e.g. for conformity assessment for the AI Act (European Commission, 2021), a developer would propose which institutions should fill the gap and how they can do so in a timely manner.

A developer provides either directly actionable and implementable measures and advice, or formulates new policy solutions to existing issues. Its political standing would be such that the solutions it puts forward have a high probability of being adopted by decision-makers, setting it aside from other institutions such as a variety of think-tanks.

The developer role brings an interesting power with it in that it can become a creator of novel policy measures, economic and financial decisions and legal proposals. The development process itself might include multiple steps, for example from the identification of a pressing policy issue to the proposed solution.

For example, the AI Act (European Commission, 2021) was developed and proposed within one institution within the EU, the European Commission. The development of a regulatory framework was achieved within the same institution that explored risk matrices, case studies and that supported the development of principled guidelines (AI HLEG, 2019), which all resulted in crucial information and suggestions included within the final proposal, the AI Act.

The type of institution proposed in this paper is envisioned more independent than this. While ensuring continuity and coherency throughout the development process of a given policy is important, it appears equally important to have some degree of political independence to ensure that the suggestions made are not only wanted but also needed.

What models of developer organizations exist? Traditionally, the development of policy proposals or the development of actionable and implementable solutions is shared between governmental agencies, departments and senior government officials such as in the UK context (Waller, 2009), or in the EU context between the European Commission, other relevant EU institutions where applicable, and member states (Wallace et al., 2020).

As sketched in this paper, however, a developer would almost serve as an external circuit breaker to ensure that the work undertaken by governments is timely, comprehensive and sufficiently in depth. In that sense, it might take the role of examining blindspots and proposing solutions for those by way of its own initiative, in addition to work it might be asked to undertake by various government agencies. Of course, it is important that in either case the developer would be in a position such that its work will be actioned or meaningfully acted upon once proposed.

What are relevant considerations? It is probable that it is difficult to set up a developer in a manner that makes it reasonably agile to account for what is needed in a rapidly changing policy space (Cihon, Maas, & Kemp, 2020). For one, it appears that this would require a considerable amount of foresight and time spent on thinking through various future scenarios to ensure that the institution is set up in an appropriate and adaptable manner.

Future orientedness. In order to ensure flexibility and a sufficient degree of future orientedness, it might serve the developer to have a mandate beyond election cycles and independent to political agendas. This consideration orients itself on the case of Rights for Future Generations and the difficulty to account for future-oriented policy making in governmental structures that haven't been set up for that purpose. Over the past year, several governments have attempted to set up a Commissioner for Future Generations within governments, to oversee various policy processes and ensure that the voice of future generations is at the table when crucial decisions are

made. While it proved not only difficult to effectively 'add on' this new role in government, most of these efforts failed after a short time (Jones, O'Brien, & Ryan, 2018). It is timely to consider an institution's ability to act with an eye towards a longer time horizon independent to the political agenda of the day, as AI is likely to be increasingly influential across various political areas in the near and longer-term future (Raso et al., 2018; Sharkey, 2019; Molnar, 2019; Nemitz, 2018) and will significantly disrupt many processes we have grown accustomed to as a society (Buolamwini & Gebru, 2018; Anderson et al., 2014; Russell, 2019).

How could this look like in practice? One shape this could take for a hypothetical EU AI Agency is that it forms an independent institution with regard to the European Commission and those located within their respective Member States. While its scope could vary, an interesting option would be for it to work on the proposals for new high-risk AI systems to be added to the legislative framework. It could act as a foresight body to EU institutions, independently reviewing and analyzing the landscape and, from that work, distilling noteworthy technological developments and proposing new high-risk AI systems in a timely manner. Its work could directly inform the adaptation procedure for new high-risk AI systems to be added to the legislative framework in collaboration with the European Commission, the European AI board and other relevant expert groups.

2.1.d. The investigator institution

Finally, the last role of an institution this paper sketches is that of an investigator.

What does it do? An investigator institution might track, monitor and investigate efforts or audit actors with regard to adherence to specific hard governance structures. It captures those abilities typically associated with a 'watchdog'. In short, it investigates whether or not actors such as governments, companies or

specific organizations adhere to the relevant standards, procedures or laws. In doing so, it also serves as an external motivator for relevant actors to ensure that they are behaving ethically.

It should be noted that investigatory ability (and the time, resources and efforts expended on it) would likely be predominantly warranted or needed only when a measure has crossed the threshold between soft and hard AI governance.

What models of investigator organizations exist? There are several 'watchdog' organizations at international level for example those monitoring human rights abuses such as Amnesty International[112] or the Human Rights Council.[113] A related model to the proposition of an investigator at European-level could be the European Court of Auditors[114] or the European Ombudsman[115] who both work to ensure that the EU is transparent in its workings and accountable for its actions. Citizens and organizations alike can file a complaint with the European Ombudsman against the EU's administration in cases of misconduct which will then be investigated. The European Court of auditors assesses how taxpayers money has been spent and reports this to the EU institutions and citizens.

A US-based example of what an investigator-type role could look like could be the independent and non-partisan Inspector Generals who conduct reviews within various Federal Agencies. They play a crucial role when it comes to government oversight and can conduct "audits, investigations, inspections, and evaluations"[116] into agency programs. *What are relevant considerations?*

Set up. The set up of such an investigator will largely depend on the number of

[112] See: https://www.amnesty.org.uk/.
[113] See: www.ohchr.org.
[114] See: https://www.eca.europa.eu/en/Pages/ecadefault.aspx.
[115] See: https://www.ombudsman.europa.eu/en/.
[116] See: https://fas.org/sgp/crs/misc/R45450.pdf.

topics it is expected to cover. The options could range from one big investigator institution for the whole topic of AI itself with multiple sub teams for specific aspects, adding teams as needed, to multiple smaller institutions ready to investigate one specific element (assuming a break by skillset, sector, etc) under a coordinator. Similar considerations and trade-offs apply when putting the role in an international context assuming that many hard governance mechanisms such as regulation will apply to more than one country. The benefits and trade-offs of one big impartial investigator versus multiple nationally located investigators needs to be weighed carefully.

The setting up of an investigator includes political considerations and if it is expected to function as a third party to those it investigates, such as governments, then it should remain sufficiently independent and impartial to those it investigates. This could become a balancing act, if the investigator equally needs to ensure that it has access to all relevant and up to date information to correctly fulfill its tasks. This is where investigative power comes into play.

Investigative Power. It makes a big difference to the impact of the investigator's output whether the investigator reviews information and data that is publicly available (or curated information made available upon request, relying on the goodwill of the institution that is being investigated), or whether it has the legal power to request access to all documentation, data, information, audits etc. including those kept internally. The investigative power itself can fall on a spectrum. The investigator may have a broad scope regarding one specific area, such as the correct implementation of legislative efforts, in which case specific investigative actions might need to be further justified before they can be undertaken. Or, it could be responsible for the investigation of a narrow scope, such as whether or not specific companies adhere to a legal criteria, in which case it is clearer what falls within its scope of investigatory power and what does not. Finally, it should be considered whether an investigator also holds the power to reprimand

actors that fall afoul their obligations or whether that power should be vested in another institution.

How could this look like in practice? One version of an EU AI Agency that acts as an investigator could be set up as a supranational agency encompassing all member states, yet as an independent actor to those member states. Depending on its size and capacity (technical and human) it could either act in response to requests made by individuals, groups or governments, to conduct ad hoc investigative exercises for a particular area, or have it within its remit to take independent action on a particular area or areas falling under its scope. For example, its scope might cover investigating the work of sub-contracted third party conformity assessment bodies that undertake assessments for AI systems that are parts of products entering the EU market, to ensure that they all have equally high standards. Given a pre-existing threat of fragmentation between various aligned efforts within the EU (Stix 2019), such a supranational AI Agency could support a coherent application of future legislative instruments and policies, while remaining politically independent.

2.1.e. Additional considerations

It is clear that, in the real world, the roles sketched in the previous paragraphs are unlikely to be completely independent of one another, and if they are, that would lead to a different set of problems (or benefits). Moreover, there are good surface-level arguments to be made about the benefit of mixing them, such as a centralization and therefore streamlining of processes, shaving off potential losses that result out of time and effort spent on communicating, updating and navigating between independent organizations and an avoidance of polluting the landscape with an increasing number of individual institutions (Cihon, Maas, & Kemp, 2020). On the other hand, it might be difficult for a larger institution to adapt to novel challenges in comparison to a network of independent institutions.

While the roles have been presented independently of one another, one could see the institutions sketched as forming links in a larger chain, increasing in complexity, power and impact. In the aforementioned cases several of the institutions could intersect with one another and take the form of hybrid institutions. For example, an analyzer that becomes specialized in a particular topic area could benefit from either morphing into an analyzer/developer hybrid or from being closely associated with a developer. Similarly, a developer/investigator hybrid may have the benefit that it has an exceptional understanding of what adherence to a specific measure would look like or not, given that it developed it, whereas independent investigator institutions may need to continuously source expertise from an independent developer institution.

Finally, it should be mentioned that there are a variety of further areas of competency that an institution can be built for that merit investigation, such as those of an enforcer, for example something such as the European Court of Justice. Unfortunately, this is outside the scope of this paper.

The paper now moves towards an exploration of the *geographical* considerations and *capacity*, complementing the landscape sketched this far.

2.2. Geography: Who are the members and what is the scope of jurisdiction?

It is increasingly evident that AI systems will impact society well beyond their original place of development or deployment (Brundage et al., 2018). Simply put, they do not respect national borders. Therefore, many AI governance concerns could be seen as multi-country concerns.

Some of the broad considerations for this axis are: What is the benefit or downside of a new multi-country institution? How does this fare in comparison to nationally 'restricted' institutions? Another consideration with regards to geography will depend on the type of AI-systems that are to be governed: for example, instances where cross-border infrastructure is desirable (e.g. autonomous vehicles) will need a different approach than those where AI-systems are deployed in public services of individual nations, that is reasonably 'restricted' to one country. In the former case, a multi-country institution might enable coherent testing procedures, common protocols, legislative efforts and smooth operation of AI-systems between affected nations. The remainder of this section will contribute pragmatic perspectives to existing literature on the topics and concerns (Cihon, Maas, & Kemp, 2020; Wallach & Marchant, 2018).

Access, Inclusion and Participation. Decisions made under the geographical axes will either reinforce and mirror, or reshape existing political alliances, "like-minded partnerships"[117] or governance efforts. A multi-country institution must therefore consider questions of access, inclusion and participation. As Koremenos et al. (Koremenos, Lipson, & Snidal, 2001b) outline: is access inclusive by design, restricted to specific states that share certain commonalities, regional or universal? Moreover, how should the institution (be able to) handle a shifting geopolitical landscape *or* expansion (including inside or outside pressure for expansion) ?

In any case, an institution with multiple countries will evangelize the chosen direction by those countries on an international level by virtue of amplifying the countries' vision, disseminating it and pooling resources to expand on it. This might lead to a competitive advantage for those nations within the institutions, who often are already comparably more powerful, effectively steamrolling efforts in individual

[117] See for example the European Union's International Alliance for a Human Centric Approach to AI. Available under: https://webgate.ec.europa.eu/europeaid/online-services/index.cfm?ADSSChck=1573028774017&do=publi.detPUB&debpub=04/11/2019&searchtype=AS&aoref=140361&orderbyad=Desc&nbPubliList=15&orderby=upd&page=1&userlanguage=en.

nations that are not part of this (larger) group (cf. this phenomenon in AI Ethics Principles (ÓhÉigeartaigh et al., 2020; Hagendorff, 2020; Mohamed, Png, & Isaac, 2020). For example, a new institution that holds any of the roles sketched under Section 2 which is predominantly populated by western actors could inadvertently outmaneuver concerns and efforts in other global regions by way of heightened visibility and amplification of the former's voice. This makes such an institution a partial actor on the global playing field and questions of access (and power) pressing.

Moveover, this scenario might also 'tip the balance': it could 'force' nations to join the effort despite it not being in their best interest given their particular ecosystem, where their alternative would be the role of a complete outsider to the seemingly new global network that is getting built. Finally, decisions to expand the membership to an institution (or not) could turn into political provocations (on purpose, or not).

At the same time, if several nations (in broad agreement) expect that their position towards AI governance is broadly more beneficial than that of other nations, it may be a reasonable political and future-oriented decision (from their perspective) to cooperate, coordinate and pool efforts between them through a dedicated institution. In short, if your belief was such that there is a threat to good AI governance from several increasingly powerful nations, one of your approaches may be to pool together with those nations that align with your vision of AI governance in order to not only level the playing field but to gain a competitive advantage and use it to the benefit of all. Whether or not that belief is correct or whether one way is more beneficial than that of another country (and which) is outside the scope of this paper. This might also contribute to a reinforced spill-over effect of the chosen and predominant path where more nations sign up to signal that they too wish to demonstrate adherence to certain (broadly beneficial) governance mechanisms.

Conversely, if nations or bigger institutions chose not to form a new institution, a proliferation of similar but distinct institutions could affect fragmentation of global AI governance regimes (and associated governance regimes). Distinct 'regimes' which have little mutual coordination, cooperation or shared expertise, might overlap or even clash. Beyond that, nation specific institutions may hinder, or complicate, a corresponding 'scaling' of AI governance measures in light of increasing AI capabilities in the coming years and decades (Müller and Bostrom 2016). To match this, and to be able to comprehensively navigate AI governance and make new institutions future-proof regardless of their geographical make-up it is likely that technical capabilities and expertise will play a crucial role. To that end, the next section will briefly investigate the third axis of this paper: *capacity*.

2.3. Capacity: *What* and *who* forms part?

This subsection introduces infrastructural considerations with regard to AI governance institutions. It is divided into *technical* and *human capacity*, though both stand in close relation. *Capacity* relates to the previous two axes (*purpose* and *geography)* in that the *what* and *where* of a given institution, will influence what the institution needs in terms of capacities for it to thrive, both from a technical and non-technical perspective.

This paper proposes that access to technical infrastructure could play an important role for future AI governance institutions. Governance proposals, policy suggestions and requirements could be improved and tailored to the state of the art (AI Index), possibly combatting the pacing gap (Marchant, Allenby, & Herkert, 2011), if the institution has capacity to accordingly run their own tests, measurements and map AI progress (see e.g. work by the AI Watch). This could range from access to compute (Brundage et al., 2020), increasing available data sets (European

Commission, 2021), to testing and experimentation facilities (European Commission, 2021). It can minimize bottlenecks in terms of information exchange, knowledge of opportunities and risks, and timeliness and increase speed between what is to be governed and the associated governance actions, decisions and proposals themselves. This could contribute to more agility, specificity and foresight in policy making for AI.

Moreover, access to technical infrastructure in-house can enable those who are developing proposals, investigating measures or otherwise, to 'fact check' ideas, possibilities and limitations without excessive reliance on third parties to provide this information to regulators or other key decision makers. Indeed, a total lack of access to technological infrastructure in house to e.g. verify claims about technical possibilities, can hand-off significant power or influence over governance decisions to other actors. These can end up becoming the sole source of information for policymakers as to what is and isn't possible to do with the technology, and are therefore capable of shaping governance measures (possibly to their interests). The power balance might not be leveled (and it might not need to be or desirable to completely do so)[118] with some degree of in-house ability to test, develop and trial run procedures but it could be significantly readjusted to benefit the suitable development of AI governance from the perspective of governments. Such a readjustment could even benefit technology companies as it is likely that hard governance efforts will have less of a 'knee jerk' quality and be more nuanced, implementable and timely than otherwise.

Infrastructure of all forms, including technical, needs individuals who can understand, handle and extract meaningful information from it: the human capacity.

[118] This paper does not mean to suggest that governmental institutions should become tech companies. It merely suggests that some in-house technical capacity (and direct expertise) can be useful for the purpose of good AI governance.

For future oriented and informed work on AI governance, it is vital to be able to reasonably comprehend, foresee, evaluate and measure a variety of scenarios with regards to the technology. This needs both technical and human capacity. Building up human capacity could take broadly two forms: (a) out of house capacity with either (i) a network of individual experts to draw on when needed[119], or (ii) an expert groups and external panels[120] and (b) in house capacity where you build up a sufficiently sized team with a range of relevant expertises such as technical, legal, and ethical, as well as diverse backgrounds, in the first place. Both (a) and (b) need never be considered mutually exclusive. For example, when soliciting expertise from external groups or networks, how can it be ensured that staff would be able to ask pertinent questions in the first instance? The answers to this question will correlate with (b) hiring diverse expertises and backgrounds: it will depend on the expertise with which the institution is populated, the degree to which individuals have the opportunity to keep updating their expertise, learning relevant new information, and engaging with technological progress.

Diversity for the purpose of this paper accounts for two things: a diverse range of expertise (including technical, legal, STS backgrounds and more), and a diverse range of backgrounds (including socioeconomic, ethnic, political and more). This is needed to cross-pollination ideas, solutions and recommendations that work for all within society, and to contribute to better and more appropriate governance mechanisms in light of AI's cross-sectoral nature. In light of this paper's proposal for an increase in technical capacity of AI governance institutions, diverse sets of expertise within staff members (such as backgrounds in various AI techniques, forecasting, or cybersecurity) would be advantageous to harness these new institutions' capacities. Individuals with a range of technical backgrounds could e.g.

[119] See for example the OECD's ONE AI network. Available here: https://www.oecd.ai/network-of-experts/.
[120] See for example the European Commission's High Level Expert Group and the working groups of the Global Partnership on AI. Available here: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai. www.https://www.gpai.ai/.

operate, manage and supervise testing or experimentation, those with a range of legal background, could e.g. work on the development of laws, those with a range of social science background could ensure governance efforts aligned with societal needs and so on, all within the same institution. This is not to say that any one background is superior to another, it is to say that *any single one background is insufficient* for the task at hand, devising good AI governance.

In order to ensure the staff has a sufficient degree of varying sets of expertise and diverse backgrounds, hiring processes and staff structures will matter a lot. While many institutions will be able to set up their own hiring processes, if you have a geographically diverse institution, it is likely that (founding) nations will expect a certain amount of representation with regards to staff within that institution (Turner, 2018). These structures may already be tied to the specific nations' political agendas without considerations for diverse subject expertise and could suggest that additional options such as external networks or expert groups are required.

3. Conclusion

In conclusion, this paper highlighted the importance of thinking about institution building for AI governance in more depth and provided a conceptual framework with which one can start working on this. In the axes on purpose, geography and capacity, the paper outlined both considerations and tradeoffs, and under purpose it also sketched connections to existing and future institutions. Future research directions for this topic could explore these axes for more concrete national cases or support decision making on new directions on an international level, especially as AI governance actions gain traction and clarity. There is also further scope to undertake investigation of additional axes and investigate their overlaps in more depth. Governments are under increasing pressure in light of AI development across all sectors (Whittlestone, Arulkumaran, & Crosby, 2021; Schiff et al., 2020; Butcher & Beridze, 2019) to come up with suitable and timely interventions and act

upon them. This paper provided one proposal towards ensuring that the infrastructure we build now to action policy and other governance proposals are considered and sufficiently future proof.

Bibliography

Abbott, Kenneth W., and Philipp Genschel. 2000. "(eds.). 2015. International Organizations as Orchestrators. Cambridge: Cambridge University Press. Abbott, Kenneth W., Robert O. Keohane, Andrew Moravcsik, Anne-Marie Slaughter, and Duncan Snidal. 2000. The Concept of Legalization." *International Organization* 54 (3): 401–19.

Abbott, Kenneth W., and Duncan Snidal. 2010. "International Regulation without International Government: Improving IO Performance through Orchestration." *The Review of International Organizations*. https://doi.org/10.1007/s11558-010-9092-3.

Almeida, Patricia Gomes Rêgo de, Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, and Josivania Silva Farias. 2021. "Artificial Intelligence Regulation: A Framework for Governance." *Ethics and Information Technology*. https://doi.org/10.1007/s10676-021-09593-z.

Alter, Karen J., and Kal Raustiala. 2018. "The Rise of International Regime Complexity." *Annual Review of Law and Social Science* 14 (1): 329–49.

Anderson, James M., N. Kalra, K. Stanley, P. Sorensen, C. Samaras, and T. A. Oluwatola. 2014. "Autonomous Vehicle Technology: A Guide for Policymakers (Rand Corporation, 2014)."

Baccaro, Lucio, and Valentina Mele. 2012. "Pathology of Path Dependency? The ILO and the Challenge of New Governance." *ILR Review*. https://doi.org/10.1177/001979391206500201.

Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1802.07228.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger,

Gillian Hadfield, Heidy Khlaaf, et al. 2020. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2004.07213.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. New York, NY, USA: PMLR.

Butcher, James, and Irakli Beridze. 2019. “What Is the State of Artificial Intelligence Governance Globally?” *The RUSI Journal* 164 (5-6): 88–96.

Calo, Ryan. 2017. “Artificial Intelligence Policy: A Primer and Roadmap.” *UCDL Rev.* 51: 399.

Cihon, Peter. 2019. “Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development.” *Future of Humanity Institute. University of Oxford*. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 2020. “Should Artificial Intelligence Governance Be Centralised? Design Lessons from History.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2001.03573.

Cyman, D., E. Gromova, and E. Juchnevicius. 2021. “Regulation of Artificial Intelligence in BRICS and the European Union.” *BRICS Law Journal*. https://doi.org/10.21684/2412-2343-2021-8-1-86-115.

European Commission. (2018). “Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Coordinated Plan on Artificial Intelligence” (COM/2018/795 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN.

European Commission. (2021). “Proposal for a regulation of the European

Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts" (COM/2021/206 final), Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

European Parliament. (2017). "European Parliament resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics" (2018) *(*2015/2103(INL))(2018/C 252/25), Brussels: Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051&qid=1620812299497.

European Parliament (2019). "Report on a comprehensive European industrial policy on artificial intelligence and robotics" (2018/2088(INI)), European Parliament website,
https://www.europarl.europa.eu/doceo/document/A-8-2019-0019_EN.html.

Erdélyi, Olivia J., and Judy Goldsmith. 2018. "Regulating Artificial Intelligence: Proposal for a Global Solution." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. AIES '18. New York, NY, USA: Association for Computing Machinery.

Garcia, E. 2020. "Multilateralism and Artificial Intelligence: What Role for the United Nations?" *The Global Politics of Artificial Intelligence*, 1–20.

Goodin, Robert E. 1998. *The Theory of Institutional Design*. Cambridge University Press.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts." *The Journal of Artificial Intelligence Research* 62: 729–54.

Hagendorff, Thilo. 2020. "The Ethics of Ai Ethics: An Evaluation of Guidelines." *Minds and Machines*, 1–22.

Hagerty, Alexa, and Igor Rubinov. 2019. "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence." *arXiv [cs.CY]*. arXiv.

http://arxiv.org/abs/1907.07892.

Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 2020. "Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence." *AI and Ethics*, October. https://doi.org/10.1007/s43681-020-00019-y.

Jones, Natalie, Mark O'Brien, and Thomas Ryan. 2018. "Representation of Future Generations in United Kingdom Policy-Making." *Futures* 102 (September): 153–63.

Kak, Amba. 2020. "'The Global South Is Everywhere, but Also Always Somewhere': National Policy Narratives and AI Justice." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 307–12. AIES '20. New York, NY, USA: Association for Computing Machinery.

Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001a. "Rational Design: Looking Back to Move Forward." *International Organization* 55 (4): 1051–82.
2001b. "The Rational Design of International Institutions." *International Organization* 55 (4): 761–99.

Kunz, Martina, and Seán Ó hÉigeartaigh. 2019. "Artificial Intelligence and Robotization." https://doi.org/10.2139/ssrn.3310421.

Marchant, Gary E., Braden R. Allenby, and Joseph R. Herkert. 2011. *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Springer Science & Business Media.

Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy & Technology* 33 (4): 659–84.

Molnar, Petra. 2019. "Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective." *Cambridge International Law Journal* 8 (2): 305–30.

Morin, Jean-frédéric, Hugo Dobson, Claire Peacock, Miriam Prys-Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 2019. "How Informality Can Address Emerging Issues: Making the Most of the G7." *Global Policy*.

https://doi.org/10.1111/1758-5899.12668.

Müller, Vincent C. 2020. “Ethics of Artificial Intelligence and Robotics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Müller, Vincent C., and Nick Bostrom. 2016. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion.” In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 555–72. Cham: Springer International Publishing.

Nemitz, Paul. 2018. “Constitutional Democracy and Technology in the Age of Artificial Intelligence.” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 376 (2133). https://doi.org/10.1098/rsta.2018.0089.

ÓhÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. 2020. “Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance.” *Philosophy & Technology* 33 (4): 571–93.

Pierson, Paul. 2000. “Increasing Returns, Path Dependence, and the Study of Politics.” *The American Political Science Review* 94 (2): 251–67.

Raso, Filippo A., Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. 2018. “Artificial Intelligence & Human Rights: Opportunities & Risks.” https://doi.org/10.2139/ssrn.3259344.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Sanders, Elizabeth. 2008. *Historical Institutionalism*. Oxford University Press.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. “What’s Next for AI Ethics, Policy, and Governance? A Global Overview.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58. AIES ’20. New York, NY, USA: Association for Computing Machinery.

Sharkey, Amanda. 2019. “Autonomous Weapons Systems, Killer Robots and Human Dignity.” *Ethics and Information Technology* 21 (2): 75–87.

Stahl, Bernd Carsten, Josephina Antoniou, Mark Ryan, Kevin Macnish, and Tilimbe Jiya. 2021. "Organisational Responses to the Ethical Issues of Artificial Intelligence." *AI & Society*, February. https://doi.org/10.1007/s00146-021-01148-6.

Stix, Charlotte. 2019. "A Survey of the European Union's Artificial Intelligence Ecosystem." *Leverhulme Centre for the Future of Intelligence, University of Cambridge*.

Stix, Charlotte, and Matthijs M. Maas. 2021. "Bridging the Gap: The Case for an 'Incompletely Theorized Agreement' on AI Policy." *AI and Ethics*, January. https://doi.org/10.1007/s43681-020-00037-w.

Thelen, Kathleen. 1999. "HISTORICAL INSTITUTIONALISM IN COMPARATIVE POLITICS." *Annual Review of Political Science* 2 (1): 369–404.

Turner, Jacob. 2018. *Robot Rules: Regulating Artificial Intelligence*. Springer.

Tzachor, Asaf, Jess Whittlestone, Lalitha Sundaram, and Seán Ó. hÉigeartaigh. 2020. "Artificial Intelligence in a Crisis Needs Ethics with Urgency." *Nature Machine Intelligence* 2 (7): 365–66.

Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. 2020. "Framing Governance for a Contested Emerging Technology:insights from AI Policy." *Policy and Society*. https://doi.org/10.1080/14494035.2020.1855800.

Wallace, Helen, Mark A. Pollack, Christilla Roederer-Rynning, and Alasdair R. Young. 2020. *Policy-Making in the European Union*. Oxford University Press.

Wallach, Wendell, and Gary E. Marchant. 2018. "An Agile Ethical/legal Model for the International and National Governance of AI and Robotics." *Association for the Advancement of Artificial Intelligence*. https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_77.pdf.

Waller, Peter. 2009. *Understanding the Formulation and Development of Government Policy in the Context of FOI*. Stationery Office.

Whittlestone, Jess, Kai Arulkumaran, and Matthew Crosby. 2021. "The Societal

Implications of Deep Reinforcement Learning." *The Journal of Artificial Intelligence Research* 70 (March): 1003–30 – 1003–30.

Winter, Philip Matthias, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. 2021. "Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/2103.16910.

Zeng, Yi, Enmeng Lu, and Cunqing Huangfu. 2018. "Linking Artificial Intelligence Principles." *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1812.04814.

Conclusion

Epilogue

AI is, and will likely continue to be, a deeply transformative technology with significant impacts on our societies. To build a flourishing future, we must remain alert to its current and future impact. We must think deeply and harness all opportunities, as they emerge, to ensure the best possible outcomes for the manifold path dependencies before us.

In the spirit of forward thinking, this thesis takes various perspectives to provide a range of possible answers to one underlying question: *how can governments — and governance practitioners — install meaningful safeguards and provide robust environments for beneficial AI development and deployment*? Although there are many key actors who will all play some part in the future of AI, this thesis focuses on governmental actors and the existing levers and communities they can draw on to support their contribution towards good AI governance.

The research contained herein has been conducted through the lens of AI governance, applied to a number of adjacent and, partially, synonymous fields, ranging from AI ethics, policy practice, and institution building, to community norms and normative concerns. Through these perspectives, the thesis draws together an adaptable framework to tackle the aforementioned inquiry, all the while devoting a particular focus to the role of one governmental actor: the EU.

Below, I summarize the main findings of this thesis. I conclude by sketching a range of practical implications and by proposing a number of additional research areas for future investigation.

1. Main findings

What is it that we are aiming for? The thesis begins by illustrating and exploring a range of conceptualizations of AI with an eye to AI governance, such as "good AI," "beneficial AI" and "responsible AI." In particular, it introduced and evaluated one of the core themes for European AI governance, informed by ethical considerations and built upon fundamental rights: "trustworthy AI."

It argues in favor of a clear and well-defined conceptualization that industry, governmental actors, legislators, policy practitioners and ethicists alike can follow and, by extension, evaluate an AI system against. In doing so, it demonstrates that "trustworthy AI" is currently the only term that is informed by technical research and academic scholarship and is sufficiently well-defined while remaining agile and future proof. Moreover, it provides a background for subsequent strategic elements within the EU's model for AI governance and demonstrates the EU's relative success in combining various research fields in a multifaceted approach, coherent with ethical considerations. This background sets the scene for the remainder of the thesis, subsequent normative claims and prioritizations chosen within each chapter.

Subsequently, Chapter I critically examines the term "trustworthy AI," pointing out its conceptual conflation and other shortcomings before delving into its memetic effect in international governance dialogue. A survey reviewing various international and European policy initiatives, strategy documents and legal text illustrates that the EU may currently hold a strategic advantage in shaping the international landscape. However, attention must be paid to ensure that well-intended proposals, measures and terms such as "trustworthy AI" maintain their intent. In reviewing the dissemination and memetic effect of "trustworthy AI" in international policy dialogue, it has become evident that the term has lost its crisp meaning, that it has become a suitcase term, rendering adherence to it meaningless. Worse, its diluted definition could obfuscate intentions, misdirect the

public's understanding and disallow true accountability and assessment of an AI system's impact.

This is an important finding for policy practitioners and scholars. We cannot pat ourselves on the back by referencing well sounding terms, efforts or measures; they must also have bite. This is especially important in light of the nascence of AI governance, the suggested impact of the EU's actions in this space, and the upcoming regulatory and standardization interventions on AI.

How can we trace, learn and extrapolate from existing governmental decisions in this nascent field? Without a doubt, the EU is making crucial moves when it comes to governing AI, not least with its upcoming horizontal regulation for AI. In order to properly locate the thesis within its 'real-world' environment, to make the research and assumptions timely and implementable, Chapter II focuses on the EU. It presents a comprehensive, in-depth analysis and review of key policy, investment and regulatory decisions informed by ethical considerations, human and fundamental rights. In doing so, it complements the review of the memetic impact of the EU's "trustworthy AI" on international governance from Chapter II and expands the scope of the thesis.

Reviewing the EU's role in AI governance serves two purposes. First, it demonstrates why the EU has been chosen as the main focus of this thesis. As highlighted above, the EU is currently the only governmental actor worldwide that is considering and, very likely, following a strategic approach towards beneficial AI from ethical guidelines through to binding horizontal regulation. Second, Chapter II contains the most complete research to date evidencing the EU's coherent and expansive approach to ethically informed AI governance. It makes a clear case that: (a) the EU has, in the past, ensured and encouraged ethical, trustworthy and reliable technological development, (b) it is currently doing so with AI and (c) there are indicators suggesting that it will continue to pursue the same path. Tracing

these developments illustrates the importance of societal trust, strong ecosystems, human-centricity and digital sovereignty to the EU's approach. I end Chapter II with a prognosis. I suggest that, given the EU's past actions and current direction, three elements within AI governance will become imminently relevant: AI megaprojects and lighthouses, standards and AI agencies. The third element of my prognosis informs the practical use case in the closing chapter of this thesis, in which I explore the future of institution building for AI governance. Since the writing of Chapter II, all three elements have become concrete topics of discussion and are likely to be formalized in the near future. In some instances, these projections even overlap. For example, the European Commission will set up a 'High-Level Forum on European Standardization', in order to support the European Standardisation Strategy and shape standards according to European values, bridging work on standards with establishing new AI-relevant infrastructures.[121]

Given that AI ethics principles are a primary recourse mechanism for AI governance, how can we increase their impact? There is no guaranteed outcome in AI governance and many measures may end up having no impact at all, despite initial promise. Developing ethical guidelines for AI is one of the primary measures most groups within the field — across all sectors and areas of expertise — have embarked on. But what is the result of these efforts? In chapter III, I investigate how the impact and operationalization of these ethical guidelines and AI principles could be reframed and increased to ensure that the work undertaken this far yields the intended results. I explore this topic with a review of the research and proposals of the High-Level Expert Group on Artificial Intelligence (AI HLEG), an independent expert group advising the European Commission on AI governance and developing policy proposals and ethical guidelines. This led me to propose a novel framework for the development of 'Actionable Principles for AI'.

[121] See: https://www.euractiv.com/section/digital/news/european-commission-poised-to-launch-high-level-expert-group-on-standardisation/.

I evaluate the pros and cons of various methodological elements in the development of ethical guidelines and present a range of lessons learned through the review of existing efforts, including those in the EU. This leads me to home in on three key findings that might increase the actionability and suitability of AI ethics guidelines for the purpose of governance efforts. These findings are important because, without ensuring that ongoing research from scholars within AI governance is applicable and actionable for current and future 'real-world' issues, we risk creating siloed communities and hindering beneficial progress for all. This risk is further investigated in Chapter IV. The three key findings to support the development of 'Actionable Principles for AI' are as follows: (1) preliminary landscape assessments, (2) multi-stakeholder participation and cross-sectoral feedback and (3) technical and non-technical mechanisms to support implementation and operationalizability.

How can different communities with varying beliefs, timelines and urgent projects efficiently collaborate on AI governance? Despite best intentions, if too much time is spent on intercommunity debates and not enough on rallying behind robust and common goals, the positive impact on AI governance — and, in particular, governmental actions — will most likely slow. The environment and the mechanisms in place within AI governance are heavily dependent on the communities working within the field. Therefore, community health is a crucial consideration in thinking about AI governance frameworks towards safe, ethical and beneficial AI. It increasingly appears that many groups working on the issues discussed in this thesis have diverging normative concerns and timelines, which has often led to confrontational disputes. Yet it is key that fragmentation of the field is avoided unless explicitly necessary. The experts in the field must remain unified to ensure that the importance of doing good with this technology remains core to the messages that are communicated to those outside the field. But how could this be achieved? In chapter IV, my co-author and I review a number of concerns and explore some epistemic, normative and pragmatic grounds for division. Subsequently, we argue that there are sufficient reasons for AI experts to converge

and cooperate in practice despite diverging beliefs. We suggest that the legal theory concept of an 'incompletely theorized agreement' could be used to ensure that cooperation is possible while each group's epistemic identity remains. In using this concept, we propose that on certain key issues, scholars working with near-term and long-term perspectives can converge and cooperate on selected mutually beneficial AI policy projects while maintaining divergent perspectives. Overall, we claim that this cooperation will strengthen the field. It would serve to avoid infighting, fragmentation and unnecessary faction building within an already small field. It would also help to raise the profile of overarching topics of agreement outside of the AI governance communities.

What institutions do we need to consider building to best support future AI governance measures? Many of the findings, lessons and sketched research questions within this thesis will need to be housed within institutional networks to be effective, actionable and impactful. Most of these institutions do not yet exist. And, as AI governance efforts are becoming increasingly concrete and pressing, their practical coming into force and monitoring becomes vital to ensure a holistic framework and overarching approach. It seems clear that new institutions or institutional networks will need to be established. I conclude this thesis by exploring this point, particularly given the context sketched throughout the earlier chapters.

Concretely, I develop a new set of blueprints for a range of institutions that could be set up to ensure that AI governance mechanisms, not least those explored in this thesis, can be implemented and will function smoothly. I present a number of sketches for building various institutions, and focusing on three key components that weem most imminently relevant: 'purpose', relating to the institution's overall goals and scope of work or mandate; 'geography', relating to questions of participation and the reach of jurisdiction and 'capacity', relating to the infrastructural and human make-up of the institution. Of those components, I

elaborate on 'purpose', as it is arguably the most pressing concern that ought to be at the forefront of scholarship on this topic — and on policy practitioners' minds.

Finally, the thesis closes the loop between all five chapters and perspectives explored. It concludes by investigating how the proposed blueprints and considerations could play out should the EU establish an AI Agency. Such an agency could implement strategic considerations touched upon and put forward in this thesis and house various epistemic communities within the field of AI governance.

2. Applicability of the research

The research presented in this thesis advances the state of the art of conventional AI governance discourse across varying subfields by focusing on the development of good AI governance frameworks. It has practical implications for the cross-pollination between various research communities and practitioners, be that for scholarly research or for the ongoing work of policy practitioners and governments. It also has implications for the overall robustness and actualization of the field going forwards.

The thesis is immersed within existing strategic directions of the field, proposing organizational methods, institutional reform and first-order conceptual considerations that can be practically applied and interlinked. It investigates and draws parallels between multiple existing efforts on a European and international level, explores avenues to strengthen them and highlights possible hurdles. Given the focus of this thesis on the EU, it is noteworthy that many of the proposed topics and considerations herein have since[122] become tangible items on the EU's agenda and gathered increasing scholarly attention, such as the likely Brussels Effect of the EU's effort to regulate AI. This underscores (i) the importance of pace and timing

[122] In this case, this refers to the publication dates of the individual papers this thesis is composed of.

when it comes to research within this field and (ii) the importance of the applicability of scholarly work to policy making, if scholars wish to affect ongoing actions at the governmental level.

Many of the predictions based on the research undertaken in this thesis and put forward in their individual chapters have found fertile ground since. For example, the establishment of a forward-looking AI Agency-model with a particular focus on general purpose AI systems has been proposed by Members of the European Parliament Pernando Barrena Arza and Cornelia Ernst ("Navigator Programme for General Purpose AI")[123] and correlates with the following questions: (1) How can we ensure community coherence and positive outcomes for policy while accounting for diverging timelines for AI (Chapter IV)? (2) How can we build future proof organizations with impactful, relevant and clear purpose (Chapter V)?

The funding for a distributed lighthouse for "safe and secure AI" established by a large cohort of research and academic institutes across the EU, the UK and Switzerland[124] highlights the applicability of the following research questions outlined in this thesis: (1) What type of AI systems do we want to have and how can we ensure their alignment and assessment through policy measures (Chapter I)? (2) How can experts from technical, legal, scholarly and policy backgrounds collaborate on a shared project (Chapter IV)? (3) How can we build on the EU's effort, funding and ecosystem to increase the likelihood of good AI governance (Chapter II)?

A key strength of the research presented within this thesis is that, in its implementation and scope, bridges practical efforts, academic research and concrete ongoing developments. For example, since the writing of this thesis' chapters, the EU has determined to set up a new High-Level Expert Group on European

[123] See: Amendment 2286 of the tabled amendments. See here: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.
[124] See: https://cispa.de/en/elsa.

Standardization.[125] This development directly correlates with three research questions considered in this thesis: (1) How can we ensure the work of expert groups on AI-relevant topics is implementable and informed (Chapter II)? (2) What direction is the EU heading in and how can we create sufficient and adequate resources now, rather than scramble later (Chapter III)? (3) How can we ensure that institutions and institutional networks are efficient and function robustly for their purpose (Chapter V)?

In particular, I explore the efficiency of the field in order to chart a distinct path for AI governance in the EU given the diverging perspectives and contributions from a range of actors (Chapter II). Given research results from Chapter I, we must also bear in mind the importance of a common language when discussing AI governance. How can we encourage accurate and meaningful terminology to capture clear AI governance intentions (Chapter I)? This last question is becoming increasingly relevant. Recent public discourse, disagreements and misunderstandings between those working within the broader fields of AI ethics and safety run the risk of negatively affecting robustly good policy efforts and increasing unnecessary community fragmentation, as addressed in Chapter IV.

3. Additional research questions

There is a lot of remaining ground to cover in this field – oftentimes, subject to distinct shifts upon new model capabilities.

I would look forward to a broader investigation of institution building and community organization for AI governance, with an eye to ensuring that we establish an environment that will function in the long term, including for many

[125] See: https://www.euractiv.com/section/digital/news/european-commission-poised-to-launch-high-level-expert-group-on-standardisation/.

generations to come. This will require significant foresight, be that through employing forecasting methods, horizon scanning or informed by possible AI observatories. It will likely also require a considerable rethinking of existing institutional infrastructures and a review as to whether they are fit and robust for the impacts that our future with AI will yield.

Another research product that I hope to see soon in academic discourse is an evaluation as to which tools are best-suited to govern AI. There is a vast opportunity in this field to develop technical measures, regulatory measures, standards, policies, codes of practice or third party oversight and testing. It is likely that all of these tools will play a significant role at different stages in the development of highly capable AI systems. However, open questions remain regarding which tools we should draw on and when. For example, should we put regulatory measures in place to slow down high-impact AI research and allow safety work to catch up? Should we allow labs to come to independent agreements outside of traditional government routes? Would international treaties lead to a slowing or complete avoidance of a race towards the bottom?

Another area in need of further investigation is the interaction between technical AI research and AI governance work. Although they are two distinct fields, in the future there will likely be an increased need for those working on AI governance to be able to access, quickly parse and comprehend technical developments in order to ensure scope-appropriate policy measures. This thesis did explore community dynamics (Chapter IV) and the impact of experts on policy measures (Chapter III), but it is likely that there will be an eventual crystallization of communities within the field that hold more power because of access to bleeding-edge information and technical understanding. It is also likely that these power dynamics will be undesirable and unilateralist, and this concern merits research attention as to how such dynamics can be mitigated.

Finally, while many open and exciting research questions remain, the economic impact — the impact of employment, workers and the potential reinforcement for existing societal inequalities — deserve their own independent research strands within the field. I hope that scholars working on these topics receive the support their research deserves. The same applies to geographical scope. Although this thesis focuses on the EU, the EU is not, of course, the sole actor engaging with this topic. Neither is the global west. The access to these technologies, as well as the global impact on all communities, deserves thorough examination to ensure that AI does benefit all, now and in the future.

Curriculum Vitae

Charlotte Stix is an AI policy researcher at OpenAI and a PhD Student fellow at the Leverhulme Center for the Future of Intelligence, University of Cambridge. She started her PhD in 2019 in the Ethics & Engineering Sciences Department at the Eindhoven University of Technology, of which the results are presented in this dissertation.

During her PhD, she worked closely with ongoing AI governance and policy processes, most notably as Coordinator of the European Commission's High Level Expert Group on AI (AI HLEG) and by leading the European AI file at OpenAI. Additionally, Charlotte has been advising governments, leading industry players and international organizations on their AI regulatory strategies, served as grantmaker to the European AI Fund and as a Fellow to the World Economic Forum's AI Council. In her spare time, she is running the widely popular EuropeanAI newsletter to better explain to civil society how, why and which decisions are made in the EU when it comes to AI.

She has been honored as a Forbes 30u30 for her contribution to the field.

List of Publications

Stix, C., Maas, M.M. (2021). Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy. *AI Ethics* 1, 261–271. https://doi.org/10.1007/s43681-020-00037-w. (in this thesis, Chapter IV)

Stix, C. (2021). Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics* 27, 15. https://doi.org/10.1007/s11948-020-00277-3. (in this thesis, Chapter III)

Stix, C. (2022). Foundations for the future: institution building for the purpose of artificial intelligence governance. *AI Ethics* 2, 463–476. https://doi.org/10.1007/s43681-021-00093-w. (in this thesis, Chapter V)

Stix, C. (2022). Artificial Intelligence by any other name: A brief history of the conceptualization of "Trustworthy Artificial Intelligence". Collection on National AI Strategies, *Discover Artificial Intelligence* 2, 26, https://doi.org/10.1007/s44163-022-00041-5. Springer. (in this thesis, Chapter I)

Book Chapters

Stix, C. (2022). The Ghost of Artificial Intelligence Governance past, present and future: AI governance in the European Union. Justin Bullock & Valerie Hudson (eds.). *Oxford University Press Handbook on AI Governance, Section 9: International Politics and AI Governance.* Oxford University Press. (in this thesis, Chapter II)

Stix, C. (2020). Zur Zulässigkeit kognitiver oder physiologischer Optimierung menschlicher Leistungsfähigkeit ohne medizinische Notwendigkeit aus ethischer

Perspektive. Hengstschläger Markus (eds.). *Digitaler Wandel und Ethik, Rat für Forschung und Technologieentwicklung*.

Technical Reports

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G.K., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensbold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J.B., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., Haas, S.D., Johnson, M.L., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A.J., Krueger, D., **Stix, C.**, Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., h'Eigeartaigh, S.'., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T.K., Dyer, L., Khan, S., Bengio, Y., & Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. ArXiv, abs/2004.07213.

Non-Technical Reports

Stix, C. (2020). The Third Way: The EU's Approach to AI Governance. In *AI Governance in 2020: A Year in Review*. Shanghai Institute for Science of Science. https://www.aigovernancereview.com/ .

Stix, C. (2019). The European Union's direction towards "trustworthy AI". In *AI Governance in 2019: A Year in Review*. Shanghai Institute for Science of Science.