National Tech Rhetoric in a Global AI Race.

Smart Futures, Public Goods and Fierce Geopolitics.

Zur Erlangung des akademischen Grades eines
DOKTORS DER PHILOSOPHIE (Dr. phil.)

von der KIT-Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

Jascha Bareis

KIT - Dekan: Prof. Dr. Michael Mäs

1. Gutachter/Gutachterin: PD. Dr. Andreas Lösch
2. Gutachter/Gutachterin: Prof. Dr. Marcus Popplow

Tag der mündlichen Prüfung: 20.03.2025

The state spends large amounts of money to enable science to pass itself off as an epic: the State's own credibility is based on that epic, which it uses to obtain the public consent its decision makers need.

Jean-François Lyotard, 1979, in the Postmodern Condition.

In memoriam:

of Milada Malinsky (†2014), Karlo Malinsky (†1976), Wilhelm Bareis (†2013), Marianne Bareis († 2013) and Maya Wanner (†2024).

As we carry our ancestors always with us in the incomprehensible myriad of ways they shape our past, present - and future to come.

Abstract

How to integrate AI technologies in the functioning and structures of our society has become a concern of contemporary politics and public debates. The dissertation investigates civil and military national AI strivings as a particular form of co-shaping this development, a hybrid of policy and discourse that offers imaginaries, allocates resources, and sets rules. Current research focuses on industry, academic, or public debates in the discursive construction of AI. Certainly, governments are impacted by public and private narratives, but, in turn, they are themselves powerful players in shaping our perception and expectation of AI. The papers of this dissertation analyze governmental positioning on AI and its role in future imaginary production, which not only includes categories of economic prosperity but similarly public good narratives, geopolitical security strivings and tensions between fact and fiction. With governments proclaiming an international AI race, they endow their imaginary pathways with massive resources and investments and contribute to co-producing the installment of these futures. Conceptually the thesis is informed by sociotechnical imaginaries, debates in technology assessment, trust, international relations, the sociology of expectations, and further, literature about the technological sublime and myths. I qualitatively analyze and compare civil (AI strategy papers) and military (position papers on Autonomous Weapon Systems) policy documents of leading AI nations of the US, China, France and Germany (selectively) towards their imaginary production of social, economic, normative and geopolitical strivings. The results of the case studies point to a reassessment of the theoretical premises of anticipating futures. It is only inadequately encircled with concepts like “vision”, “prediction”, “imaginary”, or “forecast”. A deeper understanding of future capture through hype is necessary. Conclusively, the dissertation argues that hype is a neglected concept in the study of anticipatory practices at the intersections of innovation, policy and society.

Abstrakt

Die Frage, wie KI-Technologien in die Funktionsweise und Strukturen unserer Gesellschaft integriert werden können, ist zu einem Anliegen der zeitgenössischen Politik und der öffentlichen Debatten geworden. Die Dissertation untersucht die zivilen und militärischen nationalen KI-Bestrebungen als eine besondere Form der Mitgestaltung dieser Entwicklung. Die Form dieser Bestrebungen umfasst eine hybride Mischform aus Politik und Diskurs, die der Gesellschaft Vorstellungen anbietet, Ressourcen zuweist und Regeln festlegt. Die aktuelle Forschung konzentriert sich vor allem auf industrielle, akademische oder öffentliche Debatten über die diskursive Konstruktion von KI. Sicherlich werden Regierungen durch öffentliche und private Narrative beeinflusst, aber sie sind ihrerseits mächtige Akteure bei der Gestaltung unserer Wahrnehmung und Erwartung von KI. Die Beiträge dieser Dissertation analysieren die Positionierung von Regierungen zur KI und ihre Rolle in der Produktion von Zukunftsvorstellungen, die eben nicht nur von Kategorien des wirtschaftlichen Wohlstands, sondern auch von Erzählungen über das Gemeinwohl, geopolitische Sicherheitsbestrebungen und Spannungen zwischen Fakt und Fiktion geprägt sind. Indem Regierungen einen internationalen KI-Wettlauf ausrufen, statten sie ihre imaginären Pfade mit massiven Ressourcen und Investitionen aus und tragen dazu bei, die Installation dieser Zukünfte zu ko-produzieren. Konzeptionell wird die Arbeit von soziotechnischen Imaginationen, Debatten der Technikfolgenabschätzung, Vertrauen, internationale Beziehungen, der Soziologie der Erwartungen und Literatur über das technologische Sublime und Mythos geprägt. Ich analysiere und vergleiche qualitativ zivile (KI-Strategiepapiere) und militärische (Positionspapiere zu Autonomen Waffensystemen) Politikdokumente der führenden KI-Nationen USA, China, Frankreich und Deutschland (selektiv) im Hinblick auf ihre imaginäre Produktion von sozialen, ökonomischen, normativen und geopolitischen Bestrebungen. Die Ergebnisse der Fallstudien verweisen auf eine Neubewertung der theoretischen Prämissen der Antizipation von Zukünften. Sie ist mit Begriffen wie "Vision", "Vorhersage", "Imaginäres" oder "Prognose" nur unzureichend umschrieben. Ein tieferes Verständnis von Zukunftserfassung durch Hype ist notwendig. Abschließend argumentiert die Dissertation, dass Hype ein vernachlässigtes Konzept in der Untersuchung antizipatorischer Praktiken an den Schnittstellen von Innovation, Politik und Gesellschaft darstellt.

INTRODUCTION

In recent years, national artificial intelligence (AI) strategies and regulatory initiatives have been published all around the globe and position papers have been submitted by states to the United Nations Convention on Certain Conventional Weapons (CCW). They identify potentials, risks and ethical challenges that go along AI development and its materializations such as 'intelligent' or so called 'autonomous' weapon systems (AWS). As AI seems to penetrate all spheres of life, governments are on the spot as regulators, articulating potentials, risks, and ethical challenges that go along with current AI developments. These documents do more than merely set rules: they constitute a powerful and peculiar hybrid of policy and discourse.

This thesis portrays a comparative qualitative analysis of national AI strategy papers in order to unravel these visions and to deconstruct different idealizations of statehood and algorithmic culture. States employ at the same time a prose of sober tech-policy, fierce national strategic positioning, and sketch bold visions of public goods and social order enabled through AI. Hence, policy makers are not only AI regulators but also epic storytellers, who further fire the societal imagination of AI. They become both victims and perpetrators in the AI hype production, provoking a new societal and political culture through harnessing AI with their bold policy proposals and great promises.

Firstly, this dissertation conducts a hermeneutical analysis of AI strategies from a selective choice of nations. Foremost, the state discourses around AI are theorized as revealing *Kommunikate* (empty signifiers), informing us about distinctive projections of national, cultural and institutional peculiarities. The interpretive and diffuse narrative freedom around the topic of AI allows the proclamation of sociotechnical imaginaries (SI) (Beckert 2016; Jasanoff 2015; Mosco 2005) by political leaders. As states bargain for voters' support and position themselves in the international competitive realm, they offer transformative future promises enabled through AI development. These narratives appear to serve as a means to look into a publicly communicated wished-for future and inform about societal strivings and bold aspirations in order to elicit public consent.

However, instead of being always mobilized as clear and univocal imaginaries or visions, AI in this dissertation is approached as a constant dynamic between fact and fiction. All of the

publications in this cumulative dissertation indicate that understandings and representations in multiple arenas of AI are influenced by popular culture and are inspired by more general assumptions about future technological development and its relationships to the human and society in the broadest sense. The discourses echo hopes and fears with AI substituting humans with machines, the implementation of smarter cities, real-time connected infrastructures, or the luring risks of intelligent (military) machinery that is no longer subjected to human control. These realities are thus shaped by a mix of intentional framing and larger socio-cultural narratives that act on a discursive rather than on an individual level and transcend the attribution of intentionality. Some understandings and practices of AI become socially constitutive in a respective historical period and frame the societal overall discussion.

Hence*, secondly*, the rationale of the dissertation reads that the AI realities in question can only be understood by acknowledging the constant and complex dynamic between the actual technological developments and the social histories, visions, and scenarios that are associated with them. Looking at the upcoming case studies informs us that the interpretative freedom around AI fuels different national SIs, which leads to competing imaginary language games around the phenomenon of AI. It forms worrying divergences in understanding and interpretation among political decision makers around the globe. The present ontological crisis over the *Wesenszustand* of AI becomes a negotiating crisis when arriving at conclusions about AI definitions and common (international) ethical standards and legal regulations seem no longer possible. Then, as Suchman (2023) comments "AI is invoked as a singular and autonomous agent outpacing the capacity of policy makers and the public to grasp 'its' implications. But reiterating the power of AI to further a call to respond contributes to the over-representation of AI's existence as an autonomous entity and unequivocal fact" (p.4). This process of transforming AI into a run-away force that one has to accept, implying either way playing by the rules and welcome, harness and foster its roll-out - or get drowned by the waves of technological progress - can be clearly observed in state positioning in the international realm. This rhetoric motif is especially present regarding military AI, represented by the member state's negotiations on AWS at the CCW in Geneva. Inside the national position papers submitted to the CCW circulate a variety of understandings of AI, which leads to the absurd situation that in the current debate, member states cannot even agree on whether the phenomenon of military AI, embodied through AWS, already exists or only represents

futuristic fiction. Unfortunately, here too, a mystification of AI is undertaken by political experts, as exemplified by the German AWS definition (2018) brought forward at the UN ["having the ability to learn and develop self-awareness [...] constitutes an indispensable attribute (...) to define (...) weapon systems as autonomous"]. The persistent anthropomorphization of these entities through various conceptions of *autonomy*, *consciousness*, and *intelligence* create myths and expectations of (exceeding) human performativity, which renders a fair discussion of the legal, geopolitical and normative consequences of the use of AI significantly more difficult. The Russian delegates already lamented at the start of the discussions that the process at the CCW has been of "speculative discussions divorced from reality owing to the lack of both actually operating LAWS [lethal autonomous weapon systems] and general understanding with regard to their working definition" (Russian Federation 2018). On the one hand, these linguistic confusions seem to be partly due to a simple lack of understanding of the present hyped phenomenon of AI among political decision makers. On the other hand, they seem to reflect the economic and geopolitical strategic aspirations of the respective states.

Thirdly, this deconstructed state of affairs challenges current theoretical approaches and normative foundations present in Technology Assessment (TA) policy advising and Science and Technology Studies (STS). How shall researchers and advisors react to national strivings of advancing national interest? How is public trust in AI regulatory processes affected by these strivings? While current phenomena of depoliticization through technocratic or right-wing movements are widely discussed from a variety of normative standpoints inside the TA community (e.g. Grunwald & Saretzki 2020; Delvenne & Parotte 2019; Schröder 2019), phenomena of future capture and opportunist state behavior still remains an undertheorized issue in TA and STS literature. Especially the TA community is historically closely tied to parliamentary advising, embracing a normative stance of deliberative politics, and cherishing an ethos of a common good and a responsible research and innovation perspective (RRI) (Moser 2018). If state actors do not act in such fashion but take a rather strategic and competitive stance on futures, how can TA and STS theoretically account for such maneuvers?

Fourth, the dissertation will revisit the theoretical premises that are introduced with the case studies. AI seems to be a buzzword which triggers narratives our society tells itself. It acts as a practice of collective sense making, offering orientation and guidance by articulating tacit

assumptions of shared cultural norms and social orders. However, these narratives do not need to be fleshed out full-scale narratives of a far beyond utopia, but can also be advanced as hyped stories about a nearby better world to come. McGee, a Marxist linguist who investigated the functioning of political speeches, coined the concept of the "Ideograph" in order to decode political buzzwords such as *progress, freedom* or *equality* in political discourse. He explains:

> an ideograph is an ordinary-language term found in political discourse. It is a high-order abstraction representing commitment to a particular but equivocal and ill-defined normative goal. It warrants the use of power, (…) and guides behavior and belief into channels easily recognized by a community as acceptable and laudable (McGee, 1998, p. 97).

AI seems to fulfill exactly this role of promising a laudable future. STS, TA and RRI have reacted with future directed heuristics to deconstruct and counter what is also coined as "overpromising" in the context of innovation, such as the sociology of expectations (Borup et al 2006), vision assessment (Frey et al 2022), Sociotechnical Imaginaries (Jasanoff & Kim 2009), or forecasting (Martino 2003). However, I argue that a deeper understanding of hype as an opportunist mode of anticipatory future capture is necessary. Bold statements involving fabulous potentials and shiny prospects aim to gain attention. Hereby hypers strategically narrow down future trajectories - thereby relinquishing democratic zones of public trust, imagination, and contestation.

RESEARCH QUESTIONS AND BUILT-UP

This thesis will set out to investigate selective national AI strivings in the public and military domain, narrowed down by limiting the empirical body to national AI strategy papers of four countries and the debate at the CCW. I analyze the discursive AI construction, and discuss its political and theoretical consequences for TA and its theoretical foundations. Specifically, I set out to answer the following questions, comprising three research domains that analytically follow from each other:

Empirical & Descriptive Analysis

1. *What constitutes the SIs found in the civil and military AI strategy papers by four key players in the field, namely, France, USA, China and Germany? What kinds of different national idealizations of social life and statehood are enabled through the harnessing of AI and AWS?*

2. *How is civil and military AI understood, approached and mediated by public stakeholders? What underlying larger historical, fictional or societal narratives are negotiated in the state positioning on AI.*
3. *Notwithstanding the different AI and AWS understandings and idealizations, do we see a consistency in the narrative and argumentative construction of these imaginaries?*

Regulatory and Institutional Analysis

4. *What are the direct political effects of these different AI SIs on the current geopolitical strivings of the respective states (e.g., case study of US and Chinese positioning)? Concretely, how does it contribute to or gridlock the legal and ethical regulation of* military AI inside the legislative CCW regulatory process?
5. *Given these insights, what can we learn about the relationship of trust in AI? If state actors behave strategically and opportunistically, is (dis-)trust the right response?*

Reassessment of Theoretical Foundations

6. *To what extent do concepts like SIs and technological visions not only mirror cultural and political understandings but can also be employed as strategic national weapons of confusion and deterrence?*
7. *How should/can TA and STS react to national strivings of advancing national interests and shielding off regulatory proposals? Theoretically speaking, what futuring and anticipation concepts are required to account for such maneuvers?*

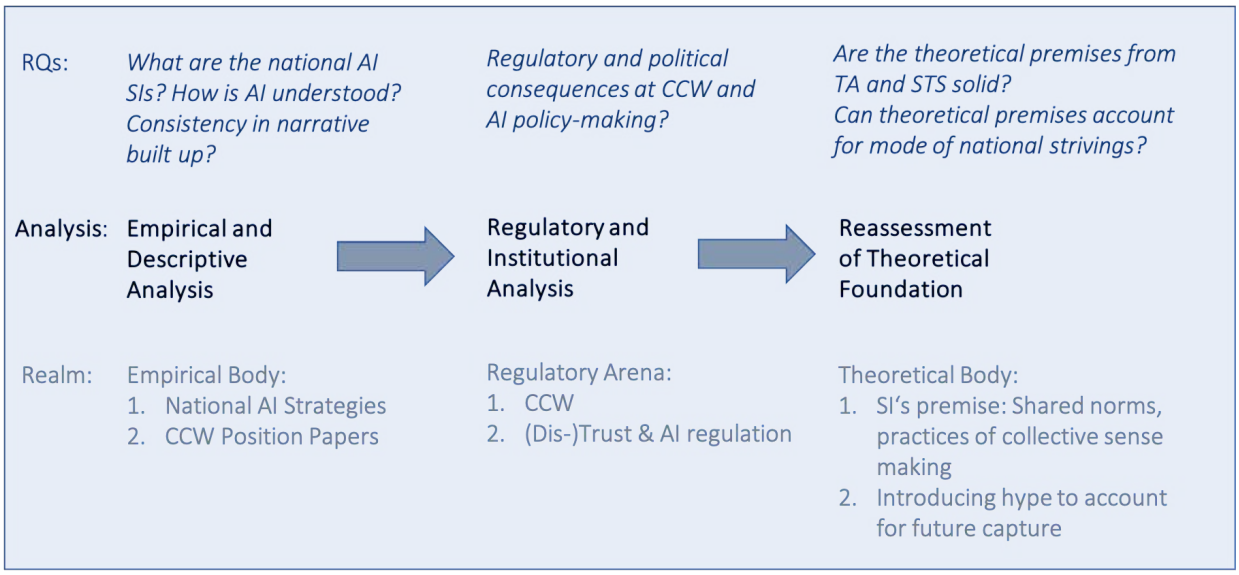


Figure 1*: Overview argumentative-built up of the dissertation*

This dissertation consists of five peer-reviewed papers (see table 1); hence, it is of a cumulative format. Each paper has its own particularities in its hermeneutical approach and theoretical take which is concretely explicated in methods and conceptual parts of the papers. These slight differences are doing justice to the different discursive arenas and societal contexts present where meaning of AI is situated, created and mediated.

Generally, the papers are unified by a constructivist understanding of AI and a hermeneutical analysis of dissecting meaning at the crossroads of fact and fiction, the real and the virtual, horror and salvation, the authentic and the fake. Hence, empirical text material is never just approached in isolation but understood as being embedded in the (often messy) social negotiations, frictions and conceptual confusions. Without a doubt, the papers show that the public and political understanding of AI lacks clarity and becomes vaguely but powerfully loaded with wishes and fears. Many debates are characterized by semantic conflations which may be (frustratingly) misguided and confusing. But these rhetoric confusions cannot be negated or corrected from a researcher perspective. To the contrary, it is exactly this space which is filled with different subjectivities and realities of their own that call for scientific interpretation, analysis and rigor. It is the endeavor of this dissertation.

Publication	Authorship	Journal	Status	Shares
Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and their Performative Politics.	Bareis, J.; Katzenbach, C.	Science, Technology & Human Values	*Published 2021*	Full execution of analysis, methods and major part in section concept. frame. In total about 75%.
The trustification of AI. Disclosing the bridging pillars that tie trust and AI together	Bareis, J.;	Big Data and Society	*Published 2024*	Everything including concept, execution of argument and theory. In total 100%.
The realities of autonomous weapons: Hedging a hybrid space of fact and fiction	Bareis, J; Bächle, T.C.	Bristol University Press	*Accepted and in printing process.*	Half of introduction, concept & section 4. Section 1, 2, 3 fully. About 65%
Lethal Autonomous Weapon Systems as a Geopolitical Signifier: US and Chinese Military Strategies as Means of Political Communication	Bächle, T.C.; Bareis, J.	European Journal of Futures Research	*Published 2022*	Major part of empirical research, military doctrines and analysis. The rest minor involvement. In total about 50 %.
Technology Hypes: Practices, Approaches and Assessments	Bareis, J.; Roßmann, M.; Bordignon, F.	Journal for Technology Assessment in Theory and Practice	*Published 2023*	*Major part of conceptualization, build up and execution.* In total about 75 %.

Table 1: *Overview of the publications*

References

Beckert, Jens. 2016. Imagined Futures. Harvard University Press.

Borup, Mads, Nik Brown, Kornelia Konrad, and Harro van Lente. 2006. 'The Sociology of Expectations in Science and Technology'. Technology Analysis & Strategic Management 18 (3–4): 285–98. https://doi.org/10.1080/09537320600777002.

Delvenne, P., & Parotte, C. (2019). Breaking the myth of neutrality: Technology Assessment has politics, Technology Assessment as politics. Technological Forecasting and Social Change, 139, 64-72.

Frey, P., Dobroć, P., Hausstein, A., Heil, R., Lösch, A., Roßmann, M., & Schneider, C. 2022.

Vision Assessment: Theoretische Reflexionen zur Erforschung soziotechnischer Zukünfte. Karlsruhe: KIT Scientific Publishing. DOI: https://doi.org/10.5445/KSP/1000142150.

Grunwald, Armin, and Saretzki, T. 2020. 'Demokratie und Technikfolgenabschätzung.

Praktische Herausforderungen und konzeptuelle Konsequenzen'.TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, 29(3), 10-55.

Jasanoff, Sheila, and Sang-Hyun Kim. 2009. 'Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea'. Minerva 47 (2): 119. https://doi.org/10.1007/s11024-009-9124-4.

Jasanoff, Sheila. 2015. 'Future Imperfect: Science, Technology, and the Imaginations of Modernity'. In Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. University of Chicago Press.

Lyotard, Jean-François. 1984. The Postmodern Condition: A Report on Knowledge. Manchester: Manchester University Press.

Martino, J. P. 2003. A review of selected recent advances in technological forecasting.

Technological forecasting and social change, 70(8), 719-733.

McGee, M. C. (1998). The "Ideograph": A Link Between Rhetoric and Ideology. In T. B. Farrell (Ed.), *Landmark Essays on Contemporary Rhetoric* (1st ed.). Mahwah, N.J: Routledge.

Mosco, Vincent. 2005. The Digital Sublime (MIT Press): Myth, Power, and Cyberspace. New Ed. Cambridge, Mass.: MIT Press.

Moser, Elias. 2018. Normative Leitbilder in der Technikfolgenabschätzung. Ita-Manu: Scripte, 18(2).

Permanent Representation of the Federal Republic of Germany to the Conference on Disarmament in Geneva. 2018. 'Statement delivered by Germany on Working Definitions of LAWS / "Definition of Systems under Consideration"', Reaching Critical Will, [online] 9 April, Available from: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Germany.pdf

Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapons Systems in the Context of the Purposes and Objectives of the Convention. 2018. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.1-WP6-English.pdf.

Schröder, Julia Valeska. 2019. 'Gegen die politische Immobilisierung durch Techno- Mythen'. TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, 28(1), 77-78.

Suchman, L. (2023) 'The uncontroversial 'thingness' of AI', Big Data & Society, 10(2): 1–5. https://doi.org/10.1177/20539517231206794

DISCUSSION: RESULTS OF THE THESIS

The upcoming section puts forward three hypotheses that can be deduced as preliminary results. These hypotheses should be understood as an overarching red-threat that structures the overall argumentation of the five papers of the dissertation and synthesizes the main take aways. Throughout the presentation and discussion of these hypotheses I will refer to the papers but also elaborate on their insights with self-standing generalizations.

1st Hypothesis: AI's Performance Is Not Only Technical But Deeply Socially Constituted

To deliver an encompassing definition of AI is a challenging task, as it remains one of the most contested and hyped concepts in the technological realm in the last decades. Even though the following technical encircling will not do justice to the scope and complexity of the term, it remains vital to problematize it and arrive at a deeper understanding of the technical functioning at hand. Notwithstanding the disciplinary perspective regarding the technical nature of AI, the proclamation of AI as a coherent research field certainly represents a rather recent historical phenomenon. The depiction of AI as a unitary discipline is overall contested, as research domains differentiated over history (Ekbia 2008). This diversity reflects the variety of existing AI approaches and research.

The technical history of AI and its paradigms

AI was first coined by computer scientist McCarthy at a 1956 conference in Dartmouth College (Russell & Norvig, 2022). The 1956 Rockefeller Foundation research proposal by the computer scientists McCarthy, Minsky, Rochester, Shannon at Darthmout aimed "to conduct a seminar on artificial intelligence (...). The study will be based on the assumption that, in principle, all aspects of learning and other features of intelligence can be described so precisely that a machine can be built to simulate these processes." In this decade scientists stemming from the fields of electrical engineering, cognitive science, information sciences, mathematics or

cryptography tried to build algorithmic machines (digital computers[1]) that could solve a given task or problem by rational and logical reasoning. They built mechanical automata that could interact with the physical environment and follow cognitive decision-making processes. As Russel & Norvig (2022) point out in their constantly updated study book on AI, "the key challenge for AI is to find out how to write programs that, to the extent possible, produce rational behavior from a small amount of code rather than from a large number of table entries" (p.45). Another recent definition of an intelligent agent reads as "any device that perceives its environment and takes action that maximises its chance of successfully achieving its goals'" (Poole, Goebel, & Mackworth 1998, p.1). The aspect of maximization and efficiency is also emphasized by Portoraro (2014), who argues that to build an automated reasoning program, one must provide a clear and concise algorithmic description to a formal calculus, allowing for its efficient implementation. As the conceptual terminology clearly shows, the *thinking* and *acting rationally* approach in AI draws heavily on other subfields such as logical formalization in mathematics, maximization theory in economics or control theory in cybernetics.

For example, utility theory, widely used in microeconomics, has early on become an essential influence on AI's genesis. It was first elaborately outlined by John von Neumann, a computer scientist, economist, and mathematician, in his 1944 book 'Theory of Games and Economic Behavior' (Von Neumann & Morgenstern 2007), and has since gained popularity through game theory approaches. Game theory paradigms are still widely used today and greatly influenced model building in the early 50s – the time where modern AI sparked. Game theory rationalizes agents and their utilities in a clearly delimited playing field, aiming at finding optimal decision-making patters. This paradigm helped to delineate and program decision- making trees for algorithms to minimize the search space and behave optimally given certain constraints. Such game theory paradigms were largely based on breakthroughs in modern formalistic mathematics in the 1920s. The momentous paradigm that prevailed at that time was the axiomatic method proposed by the German mathematician Hilbert. He postulated the elimination of the reference constraint between a 'symbolic sign' and a 'real object' (Scheich 1999). The formalistic method allowed the creation of an abstract system of symbols, in which the symbols and their relationship to each other are defined purely by inherent axioms, a

[1] The Turing machine, for example, is a model of a purely mechanical entity that exists not only in digital form, but carved out of wood: https://www.youtube.com/watch?v=vo8izCKHiF0

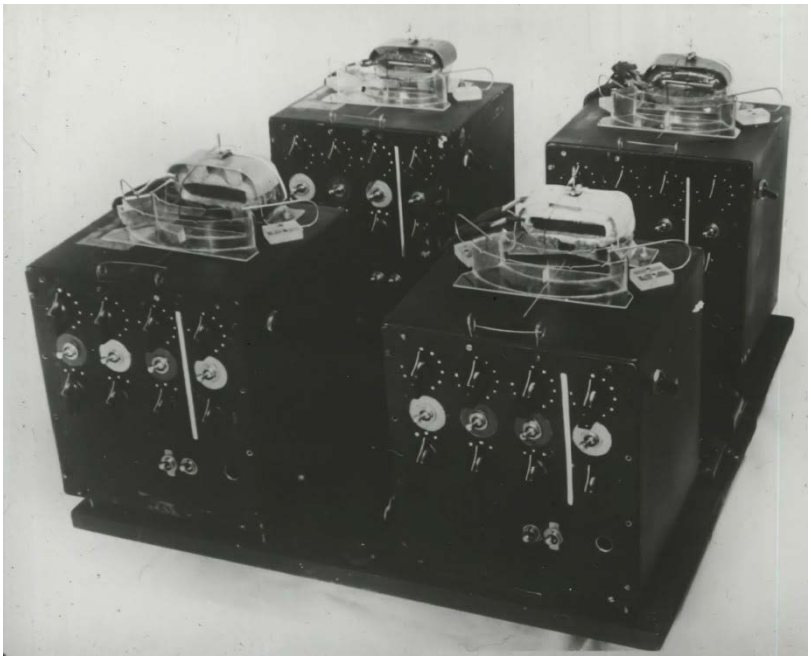
completely isolated and self-referential system that does not need to be related to the outside world (ibid.). The only requirement was the mathematical rule of consistency of the axioms. Heintz (1993) explains:

> In the formalist view of mathematics, signs and strings of signs are merely 'objects' that mean nothing, that point to nothing outside the system of which they are the building blocks. They are particles of a syntactic system that are artificially produced and mechanically processed according to the rules of the calculus (p. 57, own translation).

Even the most brilliant sequence of ideas, a complex mathematical proof, for example, can be broken down into simple operations – steps that are so simple that a machine is capable of executing them. It is these principles that characterize game theory and programming at their hearts. The symbolic approach to mathematics provided the basis for the early AI community to build formalized systems that could be solved by the sole approach of logical reasoning - computational steps that could be automated by algorithmic decision making. By setting up a closed axiomatic symbolic system, a problem or task is defined to be solved by mechanical rules. Formalisation in mathematics was reduced to a "(...) [F]ollowing rules that specify exactly how the characters may be combined and the character strings transformed" (ibid, own translation).

Sequencing is another essential foundation for the modern technical understanding of AI, as it points to the function of adaptation and learning. Instead of building simple computing machines, AI researchers needed to create rational agents that were capable to interact with the complex physical world and adapt to changes in the environment. Hence, formalization could not only be self-referential but needed to be resilient to outside inputs and shocks. As a prerequisite for adaptation, the rational agent system was theorized to have the ability to communicate and interact with the environment through feedback. Such problems of control and communication were at the core of cybernetic theory. Norbert Wiener, considered to be the founding father of the discipline, wrote in the year 1950: “The process of receiving and of using information is the process of our adjusting to the contingencies of the outer environment, and of our living effectively within that environment” (Wiener [1950] 2007, p.268). Cybernetics separates the world into the entities *system* (the rational agent) and *environment* (physical world, other agents etc.) and investigates the interaction between these entities. “Broad cybernetic philosophy [poses] that systems are defined by their abstract relations, functions, and information flows, rather than by their concrete material or components” (Heylighen &

Joslyn 2003, p. 5). Cybernetics, thus, theorizes that the goal-directed behavior of agents results from a regulatory mechanism through feedback between the system and the environment. The exchange of information allows errors and noise to be reduced in order to minimize the difference between the agent's current state and the desired goal state. Cybernetics has been crucial to the field of AI because through its system-control approach it has inspired computer scientists to create self-regulating systems that can adapt to their environment - a notion that has been sufficient for some scholars to call rational agents intelligent. For example, as Russel and Norvig (2022) write in the context of the history of AI, in reference to the cybernetician Ashby: "Ashby's *Design for a Brain* (1948, 1952) elaborated on his idea that intelligence could be created by the use of homeostatic devices containing appropriate feedback loops to achieve stable adaptive behavior" (p.15). Ashby's studies on the homeostat, which he built in the late 1940s, can be seen as one of the first electro-mechanical devices capable of adapting itself to the environment. The device would randomly reconfigure itself, influencing the polarity of the voltage. The homeostat exhibited behaviors such as habituation, reinforcement and adaption through its ability to find a stable state in a constantly changing environment (Pickering 2002).



Retrieved from: https://en.wikipedia.org/wiki/Homeostat#/media/File:W._Ross_Ashby's_1948_Homeostat.jpg

Ashby's homeostat.

As Pickering explains: "Either the device would achieve a stable configuration, in which the needle now settled at the middle of its range, or it would continue to be unstable, in which case the needle and its associated current would continue to go out of whack. (…) One could never tell from the outside how it would reconfigure itself next." (Pickering 2002, p.416 ff.)

Properties of adaptation in an unstable environment stemming from cybernetics is particularly helpful in understanding current achievements in AI based on machine learning (ML). ML applications have recently received widespread media attention, such as AlphaGo by Google's

DeepMind division, or advancements in Large Language Models (LLMs) with popular Chatbots like Chat-GPT. ML replaces the strict logical approach in AI with automated statistical decision making based on neural network modelling, which relies heavily on training data. Here, symbols within the AI model are not pre-determined, but are specified through repeated sequential trial-and-error attempts to reach a target state. As computer scientist Mahr explains, such "generation of artificial intelligence is based on a machine model that imitates the functioning and structure of the neuron network in the brain. (...) As a computer architecture, artificial neural networks give up the manipulation of meaningful symbols" (Mahr in Görz et al. 2013, own translation). Certainly, the influence of cybernetic theory on ML cannot be understanded as also Parisi (2018) notes, discussing the history of AI:

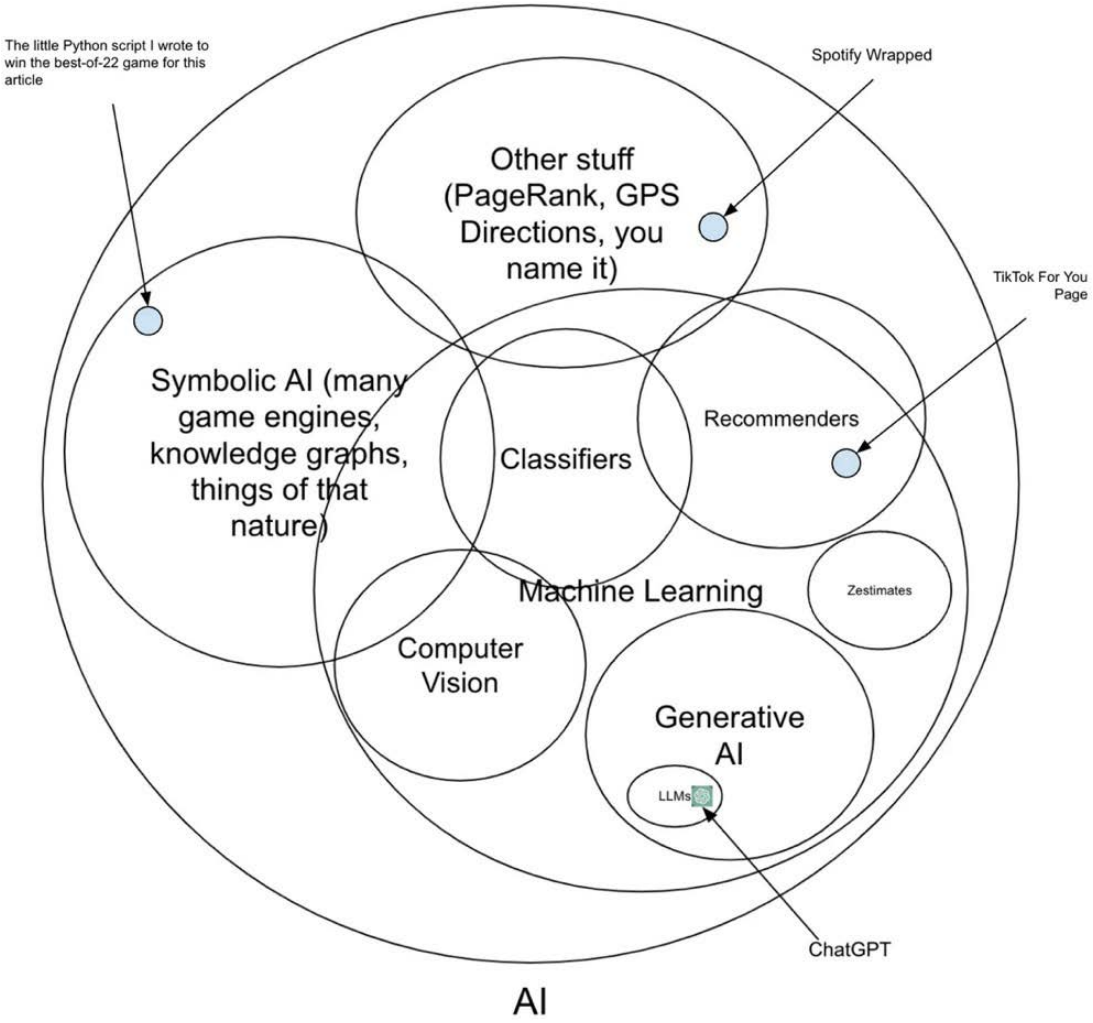
> "Early experiments in the field of cybernetics and information theory had already imagined machine learning as a form of automated intelligence that did not support a symbolic model of logical reasoning, but rather embraced a heuristic notion whereby the self-regulatory behavior of a system was supported by a trial-and-error mechanism to obtain results according to sensory data and motor skills" (p.96, own translation).

Today, writing in 2024, ML is the most discussed current state of the art in AI applications, alongside logic and knowledge-based approaches (Russel and Norvig 2022). Recent technical developments in computing power and the implementation of statistical learning theory have made this possible – but, as pointed to, the *paradigms* for these models can be found in the founding days of AI already.

There is not one AI but multiple

This genealogical tracing of AI in its historic roots reads as if AI can be understood as a monolithic phenomenon that constantly rises in performance, modelling and complexity. But AI must always be understood as multiple. There is no such thing as the 'AI' but a wide variety of models and functional types, being implemented in completely different areas of application (see illustration next page by Fraser). AIs are always expert systems in their specific domains but quickly fail if transferred into other realms. Recent public and regulatory discourses about 'general purpose AI' or 'foundation models' are in that sense misleading. Foundation refers to a process of homogenization and generalization of AI models that serve as the main building block for more specialized AI applications. The very recent policy process of the AI act on European level discussed the inclusion of such models. Art.3(1)(1)(c) of the recent AI Act proposal of the European Parliament states: "'foundation model' means an AI system model

that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks."



Taken from Fraser (2024). The clarification that there is not one AI but multiple allowing distinct tasks & applications. Accessed under: https://medium.com/@colin.fraser/generative-ai-is-a-hammer-and-no-one-knows-what-is-and-isnt-a-nail-4c7f3f0911aa

Foundation models represent a recent paradigm shift in AI (Bommasani et al 2021), particularly in speech and image processing, where they drive state-of-the-art LLM models that run popular programs such as ChatGPT, Bard, or DALL-E. They are characterized through processes of scaling and transfer learning, hence allowing to take the task learned from one application field (e.g., object recognition in images) to another task (e.g., activity recognition in social media timelines), see illustration below. However, while LLMs are powerful in a certain realm of tasks, such as analysis, structuring, captioning, or formalizing – they are surprisingly weak and unable to perform in other domains. As I have argued elsewhere (Bareis 2023), their transferability is far from universal but very limited, making the term general purpose AI a misfit.

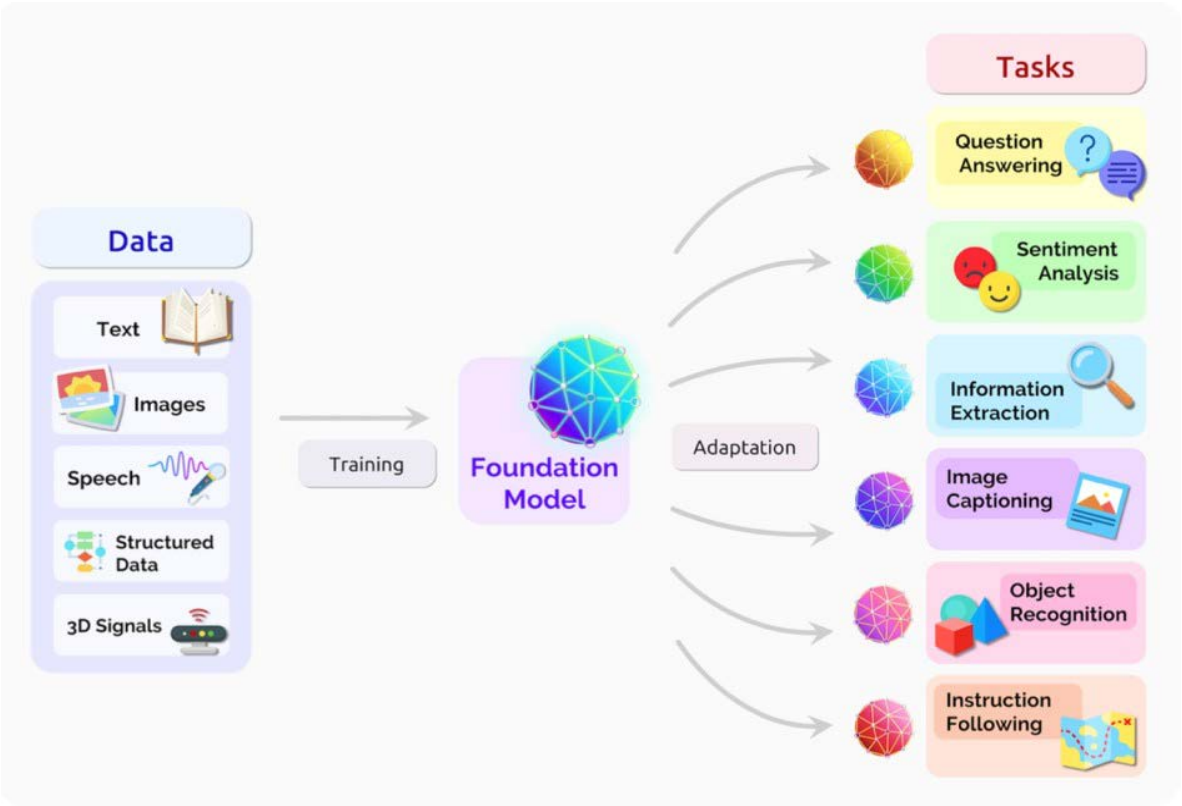


Illustration taken from Bommasani et al. 2021. Accessed under: https://arxiv.org/abs/2108.07258

For example, LLM models such as Chat-GPT fail at the simplest arithmetic tasks and calculation games, which Python programs have been able to solve for a long time, already. Fraser (2024) powerfully proves it by playing a simple 'best-of-22-game' or prompting simple counting tasks with the latest Chat-GPT model. Here Chat-GPT completely fails. This is simply because it is based on a LLM, and LLMs are probabilistic AI models. This type of model is not designed and equipped to solve exact algebra. Other AI applications such as can do this very well, but do not perform well on the syntax and rich symbolic finesse of written language. This shows that one must not speak of 'the' AI, but of different AI models and application contexts. AI models do not necessarily get better by simply feeding them with more data and computing power, even if narratives from many tech companies and linear AI history writing suggests. In public discussions, a wide variety of AI models are combined into a single amorphous entity, where the specifics and characteristics of individual models are being attributed to AI being a general agent. Fraser aptly sums up the crux of contemporary generative AI models when he writes: "The generative AI strategy is good—and getting better—at generating output that looks generally similar to examples in its training data, but it is not good at generating output that

satisfies specific criteria, and the more criteria it has to satisfy, the worse it will do" (Fraser 2024).

Anthropormophisation of AI pointing to the social and narrative force

Next to the technical conflation of AI into one phenomenon of an evolving entity, there is another historical development that accompanied the modern AI from the 1950s. The already quoted 1956 Rockefeller Foundation research proposal by the computer scientists McCarthy, Minsky, Rochester, Shannon at Darthmout College not only set out to build intelligent machines, but clarified that "the aim is to find out how machines can be made to use language, make abstractions and develop concepts, solve problems of the kind currently reserved for humans, and further improve themselves. The reference point for evaluating AI performance were humans. Hence, the phenomenon of anthropomorphization, the attribution of human characteristics onto machines, goes back to the very founding fathers of AI, above all, Alan Turing - best illustrated by the Turing test[2]. Russel and Norvig (2022) assert that the term 'intelligence' is very ambiguous and was always at risk of misleading the discussion. They argue that a more precise definition within the realm of computer science would have been appropriate, proposing that actually 'computational rationality' would be a more accurate and a less intimidating alternative to 'artificial intelligence'. But these concerns never found their place in the history of AI. Instead, its history is marked by the boldest of claims of AI surpassing human abilities, announcing a race between man (sic!) and the machine, where AI shall ultimately prevail. Computer scientists Herbert Simon proclaimed in 1965: "Machines will be capable, within 20 years, of doing any work a man can do." In 1967, Marvin Minsky hypothesized, "within a generation, the problem of creating 'artificial intelligence' will substantially be solved." Just to state three years later: "In from 3 to 8 years, we will have a machine with the general intelligence of an average human being" (Simon and Minsky in Simon 2017). Obviously, these claims did not hold, but even more striking (and worrying), these wishes and bold statements come from the *very heart* of the computer scientist expert community (Natale & Ballatore, 2017; Simon 2017). The blurring of boundaries between fact and fiction is constitutive of the historical and contemporary AI discourse, being present in many public debates (Fast & Horvitz 2017; Weber Shandwich 2016), transforming AI into a mythical and sublime endeavour. AI was historically not only ascribed the ability to imitate certain human characteristics and abilities, but also to quickly *surpass* them. Thus, I argue that the achievements and prospects of AI are

inextricably linked to modernist narratives of progress. The debate is not least a reflection of the fact that it mirrors and projects human dreams, desires and fears onto technology: futuristic epics of hopes and fears provoked by bold claims of singularities, grand narratives of superintelligences, and dreams of the automation of all tedious labor. Geraci (2008) has considered these fairy-tales as of religious qualities, embodying a messianic waiting for the singularity redemption. Some Silicon Valley engineers outspokenly follow this path, such as former Google employee Anthony Levandowski, who in 2017 established an AI church calling it 'Way of the future'. The webpage proclaims: "given that technology will "relatively soon" be able to surpass human abilities, we want to help educate people about this exciting future and prepare a smooth transition" (Way of the Future n.d.). Ekbia (2008) emphasizes the parallels between AI and a religious belief system, linking it to a modernist heritage:

> It [AI] embodies, in the most visible shape, the modernist dream of a purified world: of a mind detached from the body and the world, of a nature detached from the self and society, and of a science detached from power and politics. (…) I would therefore like to think of AI as the embodiment of a dream – a kind of dream that stimulates inquiry, drives action, and invites commitment, not necessarily an illusion or mere fantasy (p. 2).

The papers of this dissertation show that also state discourse about AI is heavily influenced by these dreams. The use of human metaphors such as artificial 'intelligence', machine 'learning' or 'autonomous' weapon systems shape the political discourse sustainably and fuels fantasies and prospective societal visions. National policymakers are influenced by pubic and expert AI discourses, turning them not only into AI regulators, but also into epic storytellers, further fueling the social imagination and mystification of AI. The dissertation shows that they become victims and perpetrators in the production of AI hype, provoking a new social and political culture by harnessing AI with their grand policy proposals and grand promises.

Consequently, to comprehend AI's dynamics and attention in society, one must acknowledge that it depicts a societal phenomenon which can never be understood by approaching it technically. Notions of efficiency, optimality, performativity, prediction, recommendation, which are inscribed in the AI phenomena do not carry legitimacy in themselves. As Jean-François Lyotard once wrote: "the language game of science desires its statements to be true but does not have the resources to legitimate their truth on its own"

(Lyotard, 1984, p. 28). In his analysis of the status of scientific knowledge and technology in the postmodern condition, Lyotard stresses that science and its materializations of technology are unable to transcend their particular mode of discourse in order to claim anything beyond their own sphere of competence and the rules by which their language game is played. He poses: "scientific knowledge cannot know and make known that it is the true knowledge without resorting to the other, narrative, kind of knowledge, which from its point of view is no knowledge at all" (ibid, p. 29). In other words, science and technology *needs* the social narrative to justify itself as valid, legitimate, needed, strived for. Lyotard elaborates on the societal function of narrative knowledge denoting roles and competences:

> [Narrative] successes or failures either bestow legitimacy upon social institutions (the function of myths), or represent positive or negative models (the successful or unsuccessful hero) of integration into established institutions (legends and tales). Thus the narratives allow the society in which they are told, on the one hand, to define its criteria of competence and, on the other, to evaluate according to those criteria what is performed or can be performed within it (ibid., p. 20).

Arguing with Lyotard, AI's status in society is inseparably tied to its epics and achievements, hopes and fears provoked through its bold scientific claims. I have argued elsewhere that the modernist narrative of AI is especially written as a story of mastering complexity – proving that the machine catches up in competition with the human (Bareis 2023). The current wave of AI, with its focus on LLMs, suggests that the solution to many of today's problems lies in finding the right answer to a Gordian task. Humanity's problems are conceptualized as a game of mastering complexity. Silicon Valley is deeply invested in winning complexity challenges in the guise of a competitive game. It is telling that complexity is presented as an essentially exclusive, elite domain that only the special, gifted, white males can handle - that is also why there are no AI mothers in the modern history of AI. AI breakthroughs are heralded with outrivaling the human in Turing's imitation game, chess, the quiz 'Jeopardy!', or board game 'GO' - and now with LLMs and their chatbots passing all sorts of exams for which they haven't been trained for. When humans lose in this competitive game, tech figures raise with a lot of sensationalism the next step in machine evolution, giving Silicon Valley the opportunity to legitimize its "disruptive" tech vision of society.

All of these tales and narratives show: AI is not only embedded in the cultural and the social - but it is *constituted* by it. The constant tension between fact and fiction, myth and reality, the overlap between technological paradigms and their larger societal and cultural manifestations

of their times cannot be separated. AI is not shared and understood as a clearly articulated, delimited, and external 'thing', 'model' or 'tool', but, as Suchman (2023, p.3) writes, "the reiteration of AI as a self-evident or autonomous technology is (…) work in progress". I follow this understanding and argue that its status is constantly being reworked in society, negotiated by many stakeholders (albeit some more powerful ones than others) across many discursive arenas where different audiences listen and take part. AI is never witnessed in code by its users but it is embedded and mediated in everyday materialities like commercial applications, social interfaces and gadgets. The dissertation argues that the perception and evaluation of AI applications and their performances are constantly negotiated by discourse, aestheticization, mediation, friction and hybridity. Strikingly, the social and cultural are not only a constitutive part of AI, they also render it performative. As this dissertation analyzes, AI provokes debates about potentials and risks, elicits emotions like hopes, fears, venture capital investments or even sparks dynamics like AI races. These social dynamics stem back from the leading early paradigms of AI like cybernetics or formalized mathematics and their hypotheses, being further transmitted by the bold claims and promises of the AI founding fathers and today's gospel of the Silicon Valley. The root and functioning of AI applications are deeply modernist (being a child of its times) but to understand their impact one has to acknowledge that AIs are woven into everyday realities, with AI applications mediating human relationships on social platforms, producing intimacies with recommender systems on dating apps, or social orders with BigTech. It is this embedding and mediation that transforms this technology into a social phenomenon. The everyday realities shape and are shaped by past and present presentations of AI, rendering the technical always in a constant and complex dynamic, which is situated differently across users, experts, temporalities, geographical locations, and constellations of power. The papers of this dissertation give powerful examples of this understanding.

2nd Hypothesis: The State And Its Contradictory Imperatives Are An Understudied Realm In TA

There is a recent debate in the Technology assessment (TA) community that TA does not only *inform* politics about the risks and potentials of technology (like AI), but that TA itself *is* politics. Historically, TA has its roots in Western democracies and features a founding narrative of practicing pragmatist policy advice to parliaments. While in its start it has still tried to maintain

the pretense of expertocratic neutrality and objectivity, TA ever since opened itself to a deliberative ethos embracing public and inclusive debate. The long established and lived reservation in the community to be associated with partiality, party politics, or loaded normativity had slowly been deconstructed by voices in the TA community. After all, “ignoring the question of normativity was not a problem as long as democracy, the separation of powers and a kind of socially acceptable technology design across party lines were taken for granted” (Nierling & Togersen 2019, p.11, own translation). Arguably, this implicit acceptance of the modernist truce was more prevalent in the 80s and 90s. Now with right-wing populism on the rise everywhere in democracies and the capitalist induced collateral damage of debts, inequalities and ecological climate change is taking its toll. And TA is not spared from scrutinizing its underlying politics. In current debates the assertion of neutrality has been criticized as a legitimation myth of an expertocratic TA reading (Delvenne and Parotte 2018; Schröder 2019, Torgersen 2018a), followed by allegations of a non-separability of facts and norms (Kollek 2019, p. 16), and even going so far to call normativity a “hidden fourth dimension of TA" (Torgersen 2018b, p. 21, own translation).

These allegations required TA to make its implicit normative assumptions transparent. Grunwald & Saretzki argue that TA stems from parliamentary advocacy and stay loyal to its foundation and argue that for calling participation, inclusion and reflexivity its normative basis (2020). With such Habermasian inspired participatory turn, TA not only tries to *inform* the delegates of political will in technology questions but also take a *reciprocal* position, receiving back impetus for its own assessments and analysis from the (ideally, diverse) public. In this democratic understanding TA takes the role of a facilitator and proactive mediator between the demos and elected executives - while being open for new impetus itself. Hence, TA not only develops assessments for executives as with parliamentary TA (for example, the TAB at the German Bundestag) but also stimulates public reflexivity, reduces complexity with ordering different viewpoints in society, or anticipates possible future pathways with scenario building (see figure 2 next page). Grunwald (2019, p.5) clearly postulates “TA’s obligation to transparency, inclusion, and democratic debate“, and thereby positions TA in the forefront of a deliberative understanding of democracy also as reaction to challenges from populist strivings or haunting historic ghosts of performing TA as expertocracy (see also the debate between Delvenne & Grunwald 2019). Including ever more realms into the sphere of problematization

and debate has been the consequent result of this Habermasian turn. Recent movements in TA like constructive TA (*inclusion* of the role of the social), Vision Assessment (*inclusion* of temporality with anticipating futures) or global TA (*inclusion* of different localities and their subjectivities while experiencing a global synchrony in problems, e.g., climate change) can be understood as a constant attempt of including other publics, epistemologies and methodologies to make a difference in the "real world".

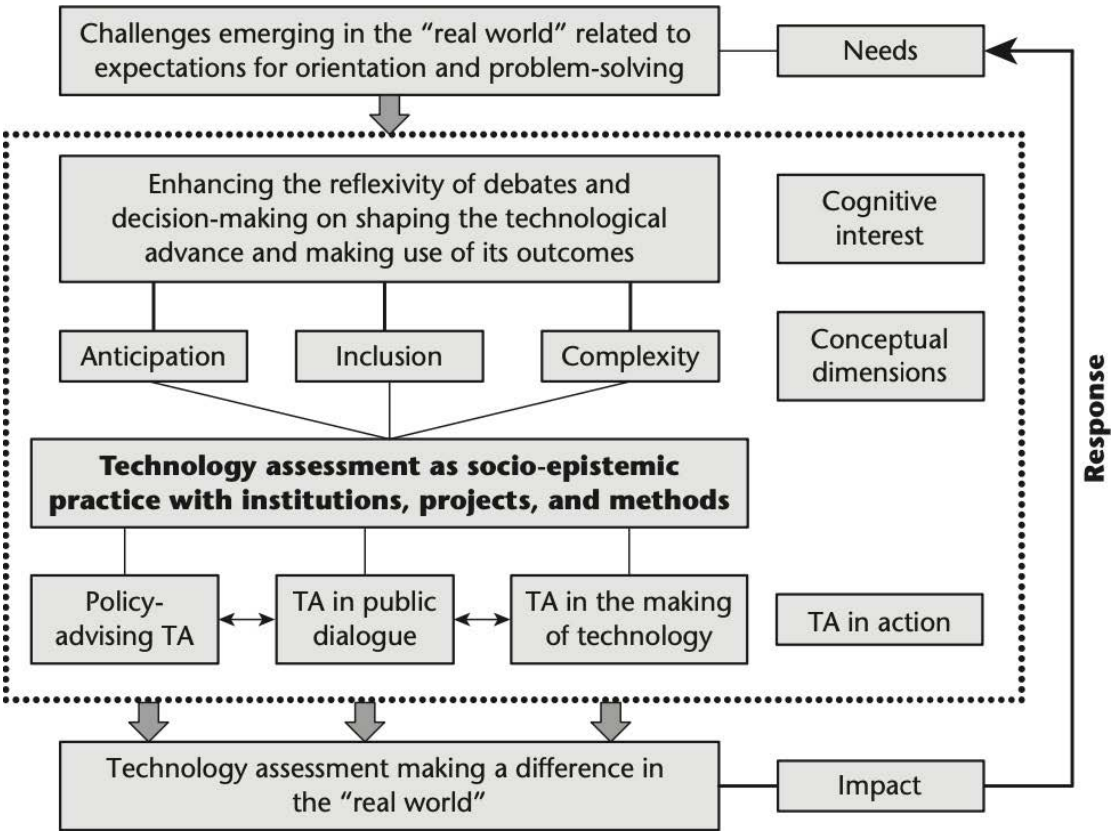


Figure 2. Grunwald's General model of Technology assessment. Taken from Grunwald 2019, p. 81

It remains striking, though, that TA's deliberative normativity is so focused on incorporating new spheres but spends little attention with the imperatives and logics that drive these spheres. The overall ideal typical and quite formalist character of deliberative politics remains abstract and ridden with prerequisites that too often clash with the messy, contradictory and irrational reality of the social and political. Much discussed in political theory, Habermas' dictum of a "forceless force of the better argument" (Habermas 1981 & 1998) and Rawls' thought experiments like the "veil of ignorance" (1971) predispose a willingness for constructive communication, openness and empathy to different subjectivities, a consideration for the greater common good, humility to subordinate oneself to the language-game of rational communication, the containment of emotions and (self-)interests and so much more. Being

situated in this deliberative ethos, parliamentary TA for example, faces many of these challenges. Just to give one example, parliamentary TA predisposes a willingness of politicians to be open to arguments concerning risks and potentials of technologies and transcend party politics - but, truly, these arguments are embedded in the realpolitik of bargaining, lobbying and political interest.

Entering AI into the picture with the case studies discussed in the papers of the dissertation illuminates the different political epistemologies, imperatives and struggles at play for TA. The

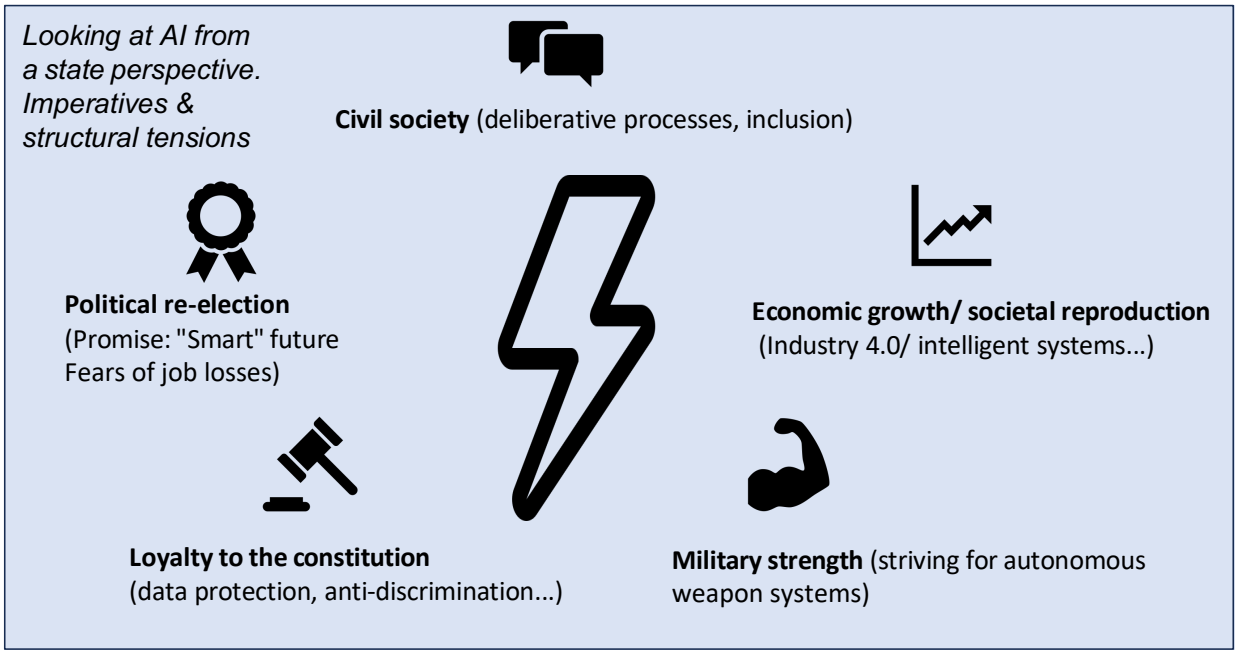


Figure 3. Theorizing the state beyond parliamentary TA.

national striving for AI indicates, so the hypothesis of this section, that the liberal state remains an undertheorized entity in TA. In TA's reading, the liberal *state* is implicitly approached as an equivalent to liberal *parliaments*. But rather than being an equivalent, it comprises just *one* part and logic of state action and functioning. There are many other pressures, language games and interests that have a say in this. Take geopolitical strivings for securing hegemonic influence and markets (see the third paper on autonomous weapon systems comparing US and China), the pressure from national corporations, trade organizations and lobbies on executives for deregulation (see the lobbying efforts from big Tech to derail the AI act as investigated in the second paper on trustworthy AI), or the personal fight of politicians and parties defending political mandate in the next election (consider the praising of executive leaders talking AI into

being in the first paper) - *all* these factors have a say in how parliamentary TA advice is received, mediated and listened to in government. Parliamentary science and technology politics is *very* much influenced by these different spheres.

In the normativity TA debate Frey et al take these pressures into account when they argue that contemporary neoliberalism practically does not allow a formalist separation of state and democracy (Frey et al 2020). They insist that the separation of economics and democracy seems incompatible with the perpetual influence of economic players on politics, for example in the interests of deregulation. They argue for a democratization of business in order to be able to achieve a democratization of technology. "The focus on economic democracy (…) broadens the perspective on what is understood by democracy: The scope of democratic values is extended to the economy, a key area of social innovation (Frey et al 2020, p.34, own translation). With the entanglement of state and capitalism they do problematize parliamentary AI in a greater context – but I would argute their take on democratizing the economy as a remedy is still falling short. This Marxist reading of TA implies that if one just *occupies* economy and state with more bottom-up democratic decision-making, the constraints on state action would vanish. But this is questionable. International pressures for competitive prices do not disappear if the decision-making process become more democratized and grass-root based.

From a more functionalist and bird-view reading of the state one has to acknowledge that this disregards many other constraints that pressure state governments and do not simply vanish by decentralization and democratization. This notion of a precarious crisis management was famously brought forward by political scientist Claus Offe (who also worked for a time together with Jürgen Habermas) in his 1973 book "*Strukturprobleme des kapitalistischen Staates. Aufsätze zur politischen Soziologie".* Offe's theory draws on the distinction of different societal subsystems, each of them following its own intrinsic logic, which he theorizes as either political (striving for legitimacy), economic (striving for reproduction) and normative (striving for re-election). With this functionalist reading of state affairs and paradigmatic turn away from the orthodox left conception based on class and class struggle, Offe became a neglected *enfant* in the traditionalist 1970s Marxist milieu. One does not have to fully agree with his functionalist, systems theory–based conception of capitalist dynamics (which was more complex anyway), but his work greatly helps to *acknowledge* that structural logics are pressuring the state, which

are in a conflicting tension to one another. Instead of *solving* conflicts with a reflexive take towards consensus as deliberative theory suggests, I would follow Offe and argue that these conflicts and contradicting imperatives cannot be solved. They can only be *handled* and *managed* in a fragile equilibrium (see a sketch of this thinking in figure 3). Politics is caught in an ensnared tension, as it has to accommodate different narratives and imperatives of interests that contradict each other in the governing of risks and benefits of technologies. Any attempt to politically solve a problem that has arisen will create new problems elsewhere. This is the idea of a precarious crisis management that is neither been accounted for in a more Marxist or deliberative reading of TA.

Wittgenstein (1953) used playing chess and its rules of the game in his philosophical investigation to give us a metaphor of how language works. One can adapt this metaphor only too well to politics. Politics is like a chessboard full of tactical moves, with the sacrifice of some chess pieces for a better overall formation, the *quid pro quo*, the pressuring of the other player's queen to let the runner free. Parliamentary TA is just another chess piece on the board of politics among many other actors.

This notion of precarious crisis management is very clearly seen in current AI policy debates. Here, for example, favouring one societal imperative like allowing ubiquitous access to user data for companies to support the rise of AI start-ups may neglect the concerns of actors pressuring governments with other imperatives. It stands in tension with users' concerns about privacy and data autonomy or enshrined constitutional ruling on informational self-determination by the German constitutional supreme court. Not only realizing but also *incorporating* this notion of politics is especially subject in the second paper about trustworthy AI, where I argue that trustworthy AI is negotiated as a constant and complex dynamic between the actual technological developments, the social realities and political power struggles associated with it. In conclusion, the publications of the dissertation tackling AI regulation illuminate and extend the understanding of normativity and politics present in TA. The tensions and contradicting imperatives in state action are neither fully grasped by a deliberative nor Marxist reading of the political in TA given the debate to date.

3rd Hypothesis: Hype Is A Neglected Conceptual Futuring Framework To Analyse Technology

The case study papers on civil and military AI and the conceptual paper on trustworthy AI urge me to revisit the theoretical and conceptual premises. What the strategy papers strikingly show is that state leaders not only inform but *show off* about their AI potency. They make the boldest of claims to reassert their national position for normative frameworks and the geopolitical arena. This "bragging" about economic gains, leadership and an efficient society enabled through AI entertains another mode of futuring than just proclaiming and inviting for a vision. The case studies show that politicians and stakeholders use all the repertoire of stage management (take hyperboles, metaphors or imageries from the rhetorical device toolbox) to build consent and legitimacy. They conceal contradictions (see the semantically empty language of trustworthy AI in the second paper), and praise themselves as competent and laudable future makers. Can future related concepual frameworks from STS and TA fully encompass the strategic nature of future capture one can witness by states rushing towards civil and military AI?

Already in 1976, Nathan Rosenberg attested a significant and neglected influence of expectations to entrepreneurial decisions and the adoption of innovation. Thirty years later, Cynthia Selin (2006) took up this note to summarize the STS' exploding interest in expectations of new and emerging technologies (NEST) in the early 2000s, centering around the sociology of expectations. Most prominent have been future directed heuristics to counter lock-in effects (Borup et al 2006), the proclamation of sociotechnical-imaginaries (Jasanoff & Kim 2015) or the undertaking of vision assessments for opening alternative trajectories (Grin & Grunwald 2000; Grunwald 2014; Lösch et al 2023). What is at stake if nations rush towards an uncontested AI future? In early 2000 Jasanoff observed that political stakeholders, experts and publics rely on technological predictions, even though the guiding visions they refer to remain incomplete, absurdly bold and lack accuracy. She noted that notwithstanding these epistemological limitations, stakeholders embrace futures with a firm determination, taking vague future talk for empirical fact. Or as Nordmann and Grunwald (2023, p.37) put it:

> Contemporary societies are obsessed with the contours of future. They do not ask what is wrong at present and how technologies can improve the situation. Instead, they witness technological change, expect more of it, and seek some kind of assurance that

> this will be a change for the better. (…) The future does not exist as yet but is nevertheless treated as an object of planning and prediction, design and contemplation.

Jasanoff warned of the "hubris" (Jasanoff 2003, p.238) and the seduction that could lock stakeholders into future promises - which could (and most probably will) turn out to be completely different. These proclaimed technological futures promised more than they could ever fulfil. In his analysis of human enhancement discourses, by similar vein, Nordmann warned of the seduction of what he called 'if and then' syndrome (2007). He identified it at the core of a new creed of speculative ethics. By means of radical "foreshortening of the conditional" speculative ethics creates forceful and unchecked futures. Speculative ethics does so by staging truisms in order to accept the obvious: "One either accepts the advent of technologically enhanced people of tomorrow or denies the obvious truth that humans have always used technology and thereby improved their condition" (Nordmann 2007, p.37). To deny the dynamics of meant-to-be change can easily result in an accusation of suffering from a pathological "status-quo-bias" or an "inappropriate favoring of the status quo", as forwarded by proponents of the ideas of longtermism (Bostrom and Ord in Nordmann 2007, p.46).

Being wary of the consequences of such forceful futures, Jasanoff called for humility and a reflective intervention in order to demask the "the normative that lurks within the technical; and to acknowledge from the start the need for plural viewpoints and collective learning" (Jasanoff 2003, p.240) instead of trusting blindly a false allure of meant-to-be future. Hence, surely, TA and STS have reacted with future directed heuristics to deconstruct and counter bold overpromising. However, the discontents raised above are mainly of epistemological nature. From this perspective, a bold future must be judged implausible, given the lack of empirical evidence and realizability. It would neither be intuitive nor rational to rely on a future trajectory that grants so little certainty. In a nutshell: Our predictions about the future can only be vague, as our knowledge is epistemologically incomplete. No one can look into the future, let alone control it.

Feminist and STS scholars would add to this epistemological limitation a social discontent. Namely, that technology, and therefore also technological futures, are always socially situated, relational and enmeshed in the messiness of society (MacKenzie & Wajcman 1999; Suchman 1987). Hence, futures are constantly reworked and understood differently by

different users and contexts. The criticism launched here is not an epistemological one, but has its roots in a social-constructivist take on the world (Bijker 1997). Subjectivities differ and so do understandings of the world and the future (take the different AI imaginaries by the states for example in the first paper of the dissertation). Objects of analysis do not have inscribed properties that carry universal understanding and truth. Hence, meaning cannot be 'found' as if it was enclosed in discrete and self-standing objects but is created in the interaction between people.

What if, however, it is exactly this epistemological incompleteness and social embeddedness that is strategically exploited by some futurists? Not because of ignorance but because of opportunism and little care about societal long-term consequences? In the coming section I argue that this mode of future anticipation through hijacking deserves more theoretical and conceptual grounding. It is only inadequately encircled with concepts like "vision", "prediction", or "forecast". Here, a deeper understanding of future as hype is necessary. The upcoming conceptual section argues that hype is a neglected concept in the study of anticipatory practices at the intersections of innovation, policy and society, further developed in the fifth paper of the dissertation.

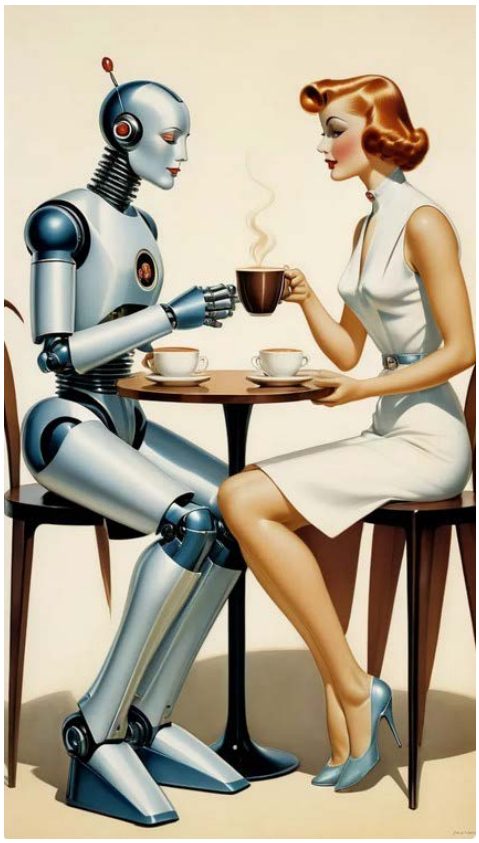
Approaching Future as Hype

Hypers must be understood as rather problematic appropriators of futures, following a deliberate mode of future capture. Instead of cultivating reflection, inclusion, alternative pathways and future literacies, hype instrumentalizes promises for the sake of creating followership and investments. Hereby hypers strategically narrow down future trajectories and relinquish democratic zones of imagination, speech and contestation. Bold statements involving fabulous potentials and shiny prospects aim to gain attention and mobilize dynamics of action. Hypers do not use plain, descriptive language to explain a coming state of events. Rather, hypers invoke bold statements and perform exaggerated claims. References to history, Zeitgeist, opportunity (costs) and risk are the common rhetorical toolkit in the creation of hype. To outsiders, this emotional hailing may seem absurd, irrational even, but it is *exactly* this emotional celebration of conquering the seemingly unachievable, or visiting the never-before-visited, which is a characteristic of hype. It is celebrating a welcomed emotional frenzy to be part of something special. There lies a deep wish in hypes to escape (even only for a moment)

everything that entangles and complicates the innocent declaration of a bold promise or opportunity at the horizon. This escapist notion takes reference to Roland Barthes' analysis of the power of myths (1972), which serve well to conceal and cope with social contradictions and help citizens to escape a sometimes-dull routine of everyday life.

Inserted prompt on AI generator 'Simplified': *"Create a society that symbolizes technology hype elicited through Artificial Intelligence*."

Generative AI depicts common stereotypes of anthropomorphism, depicting AI as a cyborgian 'other'. At the same time, it reproduces gender clichés in roles and aesthetics.



Critique in form of evidence, rationality or implausibility is often ignored by hypers. Even more, with hype it is missing the point. Hypes are sensationalist and alarming - they exploit the lack of epistemological certainty for their benefit. Hypers are masters in coining uncertainty and vagueness about the future into a story of opportunity which should not be missed. It is not that hypers lie – it is rather that they do not care about categories of truth or lie, about the difference of fact and belief. These categories are not relevant to them. In this sense hype is similar to what philosopher Frankfurt (2005) discusses with the concept 'bullshit', dissecting bullshit not as a derogative insult but developing a theoretical understanding of it as a form of communication. He writes about the relationship of truth and bullshit:

> The fact about himself that the bullshitter hides (…) is that the truth- values of his statements are of no central interest to him; what we are not to understand is that his intention is neither to report the truth nor to conceal it. This does not mean that his speech is anarchically impulsive, but that the motive guiding and controlling it is unconcerned with how the things about which he speaks truly are (p.55).

Primarily, hypes need to entertain than to talk plausibly and truthfully. This is a very indicative property of hype, because something that needs to be hailed seemingly cannot convince by its simple force of being. Hypers need to emotionally overstress the value of a future instead of the future being able to speak for itself (by plausibility, appeal, social attraction etc.). This again, though, shows the cleverness and wickedness of hypers. They do not care about truth, but they surely know how to use sensationalism to their benefit. Or again how Frankfurt would put it: In comparison to liars, bullshiters are "more expansive and independent, with more spacious opportunities for improvisation, color, and imaginative play." (p.53). And further:

> He does not care whether the things he says describe reality correctly. He just picks them out, or makes them up, to suit his purpose (…). The bullshitter is faking things. But this does not mean that he necessarily gets them wrong. In Eric Ambler's novel Dirty Story, a character named Arthur Abdel Simpson recalls advice that he received as a child from his father: Although I was only seven when my father was killed, I still remember him very well and some of the things he used to say (.) "Never tell a lie when you can bullshit your way through" (p.56 & 48f.).

What holds true for bullshit can be taken as an analogy for hype. Hype can only be understood conceptually if one acknowledges the *opportunistic* and *entertaining* character of it. These two features make the phenomenon of hype very generative. It can create topicality, produce feeds, attract investors, fuel innovation, skyrocket stocks, and give birth to start-ups, influencers, followerships of tech gurus, and other actors who know how to exploit the tech hype to their advantage. Especially, in recent years the topic received further relevance with the rise of human activity on social media platforms where people can connect and receive information on a glance (Zulli 2018), and big tech companies have perfected the attention economy.

The phenomenon of hype also connects back to the third paper of the dissertation on trustworthy AI. If users or citizens have the impression that the prospects or futures stakeholders proclaim in the public arena are not frank but strategic, it leaves a sensation of betrayal and distrust. Instead of being honest in the aims hypers pursue, they conceal real intentions like political power, the rise of market value or the attraction of investors. Seemingly, authenticity and truthfulness are one of the first virtues to burry with hypers. Also, for academic work and science communication the topic matters. Scientists are competing for funding and attention to address popular concerns, which contributes to what Vinsel (2021) and Brock &

Wangenheim (2019) recently called ‘criti-hype’, ‘ELSIfication’ and ‘wishful worries’, those terms characterizing narratives that exaggerate the perceived risks associated with a new technology, both feeding and increasing fears.

One example shall underline the theoretical from above. Bold hopes, analogies and expectations have been accompanying recent large-language data models (LLMs) and their performances since 2021. Tech firms and their leaders strategically hyped (e.g. see Gates 2023; or Future of Life Institute 2023) - and criti-hyped (Vinsel 2021) - chatbots like Chat-GPT or Bard. This leads to the worrying public misconception that the synthetical content LLMs produce can be read as factual knowledge - while early on experts warned that LLMs also produce factually wrong answers and ‘hallucinate’ (Bender et al 2021; Bommasani et al 2021). Big Tech companies wrongly hype these models as knowledge models with possibly devastating effects for society. Not only does this development show looming risks for democracies (AI Forensics & Algorithm Watch 2023), it also sheds light on the production of authority in knowledge creation. Big Tech’s and, worringly, also politicians’ framing of the LLM phenomenon powerfully informs us how speaking position in the public communication arena and impression management for the creation of followership influences trust in LLM’s synthetically created content. Non-specialist users are highly impacted by the LLM framing of providers and by the promises made by charismatic tech figures (Woznica 2022). They are seemingly trusted as truth-authorities – and, thus, also influence users’ trust relationship to technology and its output.

The concluding conceptual paper of this dissertation on technology hypes further elaborates on the practices, forms and features of hype. All in all, I hope to show that the phenomenon of hype deserves a serious conceptual place in the canon of anticipatory future concepts - instead of colloquially being treated as some mere folk talk or marketing prose.

References

Algorithm Watch & Forensic Architecture (2023). Generative AI and elections: Are chatbots a reliable source of information for voters.

Bareis, J. (2023). BigTech's Efforts to Derail the AI Act. Verfassungsblog: On Matters Constitutional.https://verfassungsblog.de/bigtechs-efforts-to-derail-the-ai-act/, DOI: 10.59704/265f1afff8b3d2df.

Bareis, J. (2023). We Are Scared of the Question Chat-GPT Cannot Answer. Because the Answer Is Too Obvious. Because the Answer Is Too Obvious.

Barthes, R. (1992). Mythologies. New York: Farram Straus and Giroux.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*(pp. 610-623).

Bijker, W. E. (1997). Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change. MIT press.

Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. Technology analysis & strategic management, 18(3-4), 285-298.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Brock, J. K.-U., & von Wangenheim, F. (2019). Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. California Management Review, 61(4), 110-134. https://doi.org/10.1177/1536504219865226

Delvenne, P., Grunwald, A. (2019). Balancing engagement and neutrality in technology assessment. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *28*(1).

Delvenne, P; Parotte, C. (2018): Breaking the myth of neutrality. Technology Assessment has politics, Technology Assessment as politics. In: Technological Forecasting and Social Change 139, S. 64–72.

Ekbia, H. R. (2008). Artificial Dreams: The Quest for Non-Biological Intelligence. Cambridge University Press.

Frankfurt, H. G. (2005). On bullshit. Princeton University Press.

Fraser, C. (2024). Generative AI is a hammer and no one knows what is and isn't a nail.

Accessed under: https://medium.com/@colin.fraser/generative-ai-is-a-hammer-and- no-one-knows-what-is-and-isnt-a-nail-4c7f3f0911aa

Frey, P., Dobroć, P., Hausstein, A., Heil, R., Lösch, A., Roßmann, M., & Schneider, C. (2022). Vision Assessment: Theoretische Reflexionen zur Erforschung soziotechnischer Zukünfte. Karlsruhe: KIT Scientific Publishing. DOI: https://doi.org/10.5445/KSP/1000142150

Frey, P., Schneider, C., & Wadephul, C. (2020). Demokratisierung von Technik ohne Wirtschaftsdemokratie? TA und die Frage demokratischer Verhältnisse in der Wirtschaft. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *29*

Future of Life Institute. (2023, March 22). Pause Giant AI Experiments: An Open Letter – Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Geraci, R. M. (2008). Apocalyptic AI: Religion and the Promise of Artificial Intelligence. Journal of the American Academy of Religion, 76(1), 138–166. https://doi.org/10.1093/jaarel/lfm101

Gates, B. (2023). The Age of AI has begun. gatesnotes.com. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Görz, G., Schneeberger, J., & Schmid, U. (2013). Handbuch der Künstlichen Intelligenz. München: De Gruyter Oldenbourg.

Grin, J., & Grunwald, A. (Eds.). (2000). Vision assessment: shaping technology in 21st century society: towards a repertoire for technology assessment. Berlin: Springer.

Grunwald, A. (2014). The hermeneutic side of responsible research and innovation. Journal of Responsible Innovation, 1(3), 274-291.

Grunwald, A. (2019). *Technology assessment in practice and theory*. Routledge.

Grunwald, A., & Saretzki, T. (2020). Demokratie und Technikfolgenabschätzung: Praktische Herausforderungen und konzeptionelle Konsequenzen. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *29*(3), 10-55.

Grunwald, A., Nordmann, A., & Sand, M. (Eds.). (2023). Hermeneutics, History, and Technology: The Call of the Future. Taylor & Francis.

Habermas, J. (1998). Faktizität und Geltung. Suhrkamp.

Habermas, J. (1981). Theorie des Kommunikativen Handelns. Suhrkamp.

Heintz, B. (1993). Die Herrschaft der Regel. Zur Grundlagengeschichte des Computers. Frankfurt am Main: Campus.

Heylighen, F., & Joslyn, C. (2003). Cybernetics and Second-Order Cybernetics. In Encyclopedia of Physical Science and Technology (pp. 155–169). https://doi.org/10.1016/B0-12-227410-5/00161-7

Jasanoff, S. (2005). Technologies of humility: Citizen participation in governing science. Minerva 41: 223-244.

Jasanoff, S., & Kim, S. H. (2015). Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power. University of Chicago Press.

Jasanoff, Sheila; Kim, Sang-Hyun (2009): Containing the atom. Sociotechnical imaginaries and nuclear power in the United States and South Korea. In: Minerva 47 (2), pp. 119–146. https://doi.org/10.1007/s11024-009-9124-4 DOI: https://doi.org/10.1007/s11024-009-9124-4

Kollek, R. (2019). Implizite Wertbezüge in der Technikfolgeabschätzung. Plädoyer für eine Praxis der reflexiven Normativität. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 28 (1), S. 15–20

Lösch, A.; Schneider, C; Dobroć, P; Frey, P; Gondolf, J; Hausstein, A ;Heil, R. (2023) Vision assessment: An orientation framework for the practice of technology assessment. KIT Scientific Working Papers. DOI: 10.5445/IR/1000158791

Lyotard, J.-F. (1984). The Postmodern Condition: A Report on Knowledge. Manchester: Manchester University Press.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. Accessed from: https://web.archive.org/web/20070826230310/http:/www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

MacKenzie, D., & Wajcman, J. (1999). The social shaping of technology. Open university press.

Nierling, L., & Torgersen, H. (2019). Normativität in der Technikfolgenabschätzung: Einleitung in das TATuP-Thema. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und*

Praxis/Journal for Technology Assessment in Theory and Practice, *28*(1), 11-14.

Nordmann, Alfred (2007): If and then. A critique of speculative NanoEthics. In: NanoEthics 1 (1), pp. 31–46. https://doi.org/10.1007/s11569-007-0007-6

Offe, C. (2006) [1972]. *Strukturprobleme des kapitalistischen Staates: Aufsätze zur politischen Soziologie*. Campus.

Parisi, L. (2018). Das Lernen lernen oder die algorithmische Entdeckung von Informationen. Machine Learning–Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz. Bielefeld: transcript, 93-114.

Poole, D. I., Goebel, R. G., & Mackworth, A. K. (1998). Computational Intelligence: A Logical Approach. New York: Oxford University Press.

Pickering, A. (2002). Cybernetics and the mangle: Ashby, Beer and Pask. Social studies of science, 32(3), 413-437.

Portoraro, F. (2014). Automated Reasoning. Stanford Encyclopedia of Philosophy. Retrieved from: https://plato.stanford.edu/archives/win2014/entries/reasoning-automated/

Rawls, J. (1971). A theory of justice. Cambridge Universtity Press.

Rosenberg, N. (1976). On technological expectations. The economic journal, 86(343), 523 535.

Russell, S., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach, 4th, Global ed. Selin, C. (2007). Expectations and the Emergence of Nanotechnology. Science, Technology, & Human Values, 32(2), 196-220.

Sharma, Abz; Grant, David (2011): Narrative, drama and charismatic leadership: The case of Apple's Steve Jobs. In: Leadership 7 (1), pp.3–26. https://doi.org/10.1177/1742715010386777

Scheich, E. (1999). Technologische Objektivität und technische Vergesellschaftung. Identitätslogik im naturwissenschaftlichen Diskurs. Zur Veränderung erkenntnistheoretischer Perspektiven durch die Konstruktion und Politisierung der Natur. In M. Ritter (Ed.), Bits und Bytes vom Apfel der Erkenntnis. Frauen-Technik-Männer. Münster: Westfälisches Dampfboot.

Schröder, J. V. (2019). Das Politische in der Technikfolgenabschätzung: Reflexionen mit der pluralen, radikalen Demokratietheorie von Laclau und Mouffe. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis/Journal for Technology Assessment in Theory and Practice*, *28*(3), 62-67.

Simon, J. (2017). Fascinating Tales of a Strange Tomorrow. Retrieved November 22, 2019, from Medium website: https://towardsdatascience.com/fascinating-tales-of-a-strange-tomorrow-72048639e754

Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge university press.

Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2), 20539517231206794.

Torgersen, H. (2018a): Three myths of neutraliy in TA. How different forms of TA imply different understandings of neutrality. In: Technological Forecasting and Social Change 139, S. 57–63.

Torgersen, H. (2018b): Die verborgene vierte Dimension. Normative Reflexion als Erweiterung der Theorie der Technikfolgeabschätzung. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 27 (1), S. 21–27

Vinsel, Lee (2021): You're doing it wrong. Notes on criticism and technology hype. In: Medium, 01.02.2021. Available online at https://sts-news.medium.com/youre-doing-

it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5
Von Neumann, J., & Morgenstern, O. (2007). Theory of Games and Economic Behavior (60th Anniversary Commemorative). Princeton, N.J. ; Woodstock: Princeton University Press.
Way of the Future. (2019). Retrieved from http://www.wayofthefuture.church/
Weber, S. (2016). AI-Ready or Not: Artificial Intelligence Here We Come! Retrieved November 26, 2019, from Weber Shandwick website: https://www.webershandwick.com/news/ai-ready-or-not-artificial-intelligence-here-we-come/
Wiener, N. (2007). Cybernetics in History [1950]. In R. T. Craig & H. L. Muller (Eds.), Theorizing Communication: Readings Across Traditions. SAGE.
Wittgenstein, L. (1953). Philosophical Investigations. Accessed only under: https://static1.squarespace.com/static/54889e73e4b0a2c1f9891289/t/564b61a4e4b04eca59c4d232/1447780772744/Ludwig.Wittgenstein.-.Philosophical.Investigations.pdf
Woznica, M (2022): Stage performances as means for linking sociotechnical imaginaries and projective genres in the discourse around urban air mobility. In: European Journal of Futures Research 10 (12). https://doi.org/10.1186/s40309-022-00198-3
Zulli, D. (2018). Capitalizing on the look: insights into the glance, attention economy, and Instagram. Critical Studies in Media Communication, 35(2), 137-150.

WALK-THROUGH THE DISSERTATION PAPERS & OVERALL RED THREAT

The five papers of this dissertation can be divided into three different parts. The first two papers address the empirical arena on civil AI (PART I), article three and four tackle the empirical arena of military AI (PART II), and the last article tackles the reassessment of the theoretical foundations (PART III). In this section I will walk through the main line of thought of the papers and their overall connection.

The first scientific paper *Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics*, was published together with Prof. Dr. Christian Katzenbach from the university of Bremen in the STS flagship Journal *Science, Technology and Human Values* in 2021. It ranks as the most cited the paper in the journal in the last three years and also the most read paper in the last three months (as of 2024).

The paper analyzes existing state narratives of a present and future society to be transformed by AI, and gives an account of regulatory initiatives in the international comparative governance framework. We empirically analyze the AI policy documents of four key players in the field, namely China, the United States, France, and Germany. From a political and economic perspective, the national AI policy documents show many similarities in order to become economic competitive leaders. The narrative construction of AI strategies is strikingly similar: all states establish AI as an inevitable and massively disruptive technological development, relying on rhetorical devices such as stressing grand legacy and international competition. On a cultural level, however, they differ fundamentally in their values, approaches, and scenarios of what the materialization of AI (e.g., Smart City or Industry 4.0) should look like. In doing so, they draw on familiar national narratives and founding myths. In order to carry out this hermeneutic analysis, AI in this publication is not understood as a constant, self-contained and purely technical concept, but as a complex dynamic between current technological developments and associated possible social futures. The discussion of AI is embedded in a broad social discourse in which national narratives of cultural identity meet a technical world between fact and fiction, wishful thinking (for example, liberation through automation of work) and social scenarios of fear (versus job losses through automation of work). Worryingly, the results of the cross-national analysis show that governments around the world contribute to talking AI into being, rather than acting as public watchdogs. They act as performative AI

hypers instead of critically assessing the risks and potentials of this controversial technology. While one would expect the formulation of a future and the subsequent question what role AI should play in it, the future projected by governments is highly mediated and constituted *by* technology, making AI not only an enabler but also a remedy for deep-seated societal problems such as inequality, urbanization, or climate change. With this publication I have established a comparative consolidation of rationales, narratives, regulatory approaches and government structures towards AI. The analysis clearly shows the risk that policy makers can fuel a hype instead of containing it. Regulatory institutions create AI governance frameworks that are exposed to a tension between opportunities and risks, expectations and fears, economic interest and protection of users. The present and future use is marked by many systemic and individual risks caused by AI and policy makers must build robust policy that can bridge these insecurities and conflicts of interests between many stakeholders. In order to (re)establish trust in their leadership and in AI, trust saliently appears as a governing rational across AI policies. The terminology of trust appears frequently in AI governance papers, without clearly stating its function in policy frameworks.

Therefore, the second publication of the dissertation engages with scrutinizing trust as an AI governance principle. Next to a rich ethical discourse on trustworthy AI, one can witness the concrete formation of a global governance regime around AI (the consequent manifestation of tech *talk* into *written* policy as announced in the publication before). This governance regime consists of an overlapping ensemble of private standards, normative principle-setting and concrete standardization efforts, which are heavily featuring notions of trust. The OECD and the AI 'EU ecosystem' favour a risk-based regulatory approach, which is branded as, citing the name of the EU AI act, "excellence and Trust in Artificial Intelligence". However, while trustworthy AI is trending high on the political agenda, the term is utterly ill- defined. What is actually meant when talking about trustworthy AI, and why it is so difficult to achieve, remains insufficiently understood by both academic discourse and current AI policy frameworks. In the paper *The Trustification of AI. Disclosing the bridging pillars that tie trust and AI together* aims to close this research gap. The paper appears 2024 in the journal *Big Data & Society*, which represents one of the leading scientific platforms on the debate on AI, where pressing social, cultural and political issues of the integration of AI into our society is critically reflected upon.

The main aim of this paper is not to assess whether AI is trustworthy or not, but to give an account of the *dimensions* that need to be considered in order to be able to assess it. Does it make sense at all to talk about trust in the latter case or are we just dealing with a conceptual misfit? After all, there exist valid philosophical reservations about simply transferring interpersonal trust to human-machine trust which are instructive for the overall trustworthy AI debate. The way trust is handled in both the policy and academic AI debate is very sloppy, staying undertheorized and just taken for granted in colloquial use. Users approach trust and AI as something intersubjective, expecting great things from their new AI powered gadget and then being utterly disappointed if it fails to do so. Users perceive AI as something being highly mediated by powerful actors, as when Elon Musk trusts that AI will be the cure to the world's problems, many people seem to follow blindly (but do they trust AI then or Elon Musk?). And as something that can mobilize greater political dimensions and strong sentiments. As when a friend of mine told me that she would certainly distrust AI because she distrusted the "corrupt" politicians who instead of regulating it, let big Tech "take the profits and get rich without taking care of the larger societal consquences". Communication, mediation, sentiments, expectations, power, misconceptions – all of this seems to have a say in the relationship between AI and trust. This creates a very messy picture with AI and trust being enmeshed in a social complex interplay with overlapping epistemic realms. To make sense of this picture the paper draws on multiple inspirations, such as phenomenology to reveal AI as a quasi-other that we (dis)trust; STS to deconstruct the social and rhetorical embedding of AI; and political science to identify hegemonic conflicts in regulatory negotiations. I come up with an analytical scheme which shall serve as a heuristic to better understand the contested phenomenon of trustworthy AI. The paper appeared in the context of the ITAS focus group, "gesellschaftliches Vertrauen in lernende Systeme" (GVLS), led by Reinhard Heil. It sums up insights already developed in the overall poster of the group.

The paper on trust is also attempting to find an answer to the issue raised by the first publication, namely, why is it that governments around the world tend to fuel the AI hype instead of regulating it? Rather than theorizing policy as a deliberative process between different actors, I argue for an agonistic picture of governance, depicting strivings for hegemony and agenda setting between players and the difficulty of deciding upon value trade-offs. Hence, the AI policy process is better understood as a *bargaining field* of conflicting actors trying to maximise their stakes. Politics is caught in a mediating tension with AI regulation, as it has to

accommodate different narratives and imperatives of interests that contradict each other. In the paper I argue that bold AI narratives have a dual political function. First and foremost, they shall trigger attention. Promising a shiny AI future shall endow executives with the aura of guiding leadership and legitimacy to build visionary futures. Simultaneously, the tech-talk shall spur financial investments in the AI market, trying to secure market shares in a global competitive market. Politically, however, the bold but vague future statements can be interpreted, and *problematically* so, as a means to sugar-coating conflicting value trade-offs and as a handy means to managing expectations through playing around with hopes and fears associated with AI. In that sense the semantic carving out of trustworthy AI in the political sphere may not only be the consequence of a scattered debate but may also depict political strategy.

The theme of deliberate political strategy is certainly the overall theme for the third publication of the thesis, changing the scene from the public AI arena to the military AI debate around autonomous weapon systems. What seems implicit and suggestive in state communication around public AI becomes very explicit with its military use for geostrategic strivings. It is very striking that domestic and European AI regulation only targets the civil use of AI and completely neglects the regulation of the military domain. The European AI act, for example, completely ignores the military application field, let alone the dual-use character of AI. However, to understand states' strivings towards AI, it is strictly essential to approach state administrations as *strategic* actors who not only regulate domestic and civil AI use but also try to secure and amplify their geopolitical interests in the international sphere. Hence, the arena the third publication looks at is the global political domain, where a world order can be witnessed where many national powers strive for hegemonial influence. In the paper *"Autonomous weapons" as a geopolitical signifier in a national power play: Analysing AI imaginaries in Chinese and US military* policies, published in 2022 in the *European Journal of Futures Research*, with Dr. Thomas Christian Bächle, head of the Digital Society Research Programme at the Alexander von Humboldt Institute for Internet and Society (HIIG) in Berlin, I analyze how AI talk becomes weaponised as tools in the geopolitical arena. The regulatory discussion around the military use of AI is taking place on United Nations (UN) level, consisting of position papers in the debate at the Convention on Certain Conventional Weapons (CCW) in Geneva. Here, the ways in which nation-states portray themselves as part of a global AI race, competing for economic, military, and political advantage, become evident. This is especially

true for China and the United States, which are seen and perceived not only as international hegemons but also as antagonists promoting competing self-understandings. This is reflected in their histories, political doctrines, and national identities. The paper zooms into this case study of these two nations. It conceptualizes AWS as a geopolitical signifier and approaches the military standpoint and strategy papers as a form of political communication that is pursued as being part of military AI imaginaries. AWS are a central element of the goals both China and US pursue in the realm of geopolitical strivings. Differing definitions and normative understandings of AWS are *deliberately* employed to serve national interests and, consequently, making it more difficult to reach a UN regulatory consensus. As of 2024 the UN regulation on military AI is gridlocked and for many commentators has essentially failed. Many states like US, China, Israel, Korea or Russia have made clear that they have an interest that military AI research and their applications will not be hindered (among other states like Germany and France, who are not really willing to take sides pro or against AWS). At the moment, the world is caught in a military AI race, being fueled by a worsening of the geopolitical situation with the invasion of Russia in Ukraine, a tightening of the Taiwan crisis or the war in the middle East. Exchanging with other commentators and researchers on this situation after the publication of the paper at conferences, the geopolitical situation seems only worryingly worsening at the moment (as of April 2024).

The subsequent fourth publication "*The realities of autonomous weapons: Hedging a hybrid space of fact and fiction*", again co-authored with Dr. Thomas Christian Bächle, further investigates how understandings and meanings around AWS are constructed as a complex entanglement. The article serves as an introduction to an edited volume, peer-reviewed and to be published by Bristol University Press (manuscript in editing and printing stage, published in early 2025), though, it is written as a research article. As the US-China comparison in the paper before shows, current political military arenas of autonomous weapons get mixed up with understandings from popular culture, regulatory debates, journalism and research. That is why the fourth article analyzes how the current debates on AWS mediate between fact and fiction and create a constant and complex dynamic between the actual technological developments and the potential futures that are associated with them. Paradoxically, it is exactly in this context of uncertainty – in which reality, imagination, possibility and fiction are conflated – that the full scope of this controversial technology becomes visible. Hence, the article focuses on various practices, discourses and techniques in which AWS are both represented and created

to become technological, military and political realities. It adheres to an analysis of the different meanings articulated across multiple domains that constitute the "realities of autonomous weapons" and powerfully influence how we perceive and engage with these novel technologies. The article puts forward five reflections in its analysis that are meant to pinpoint these complex realities of autonomous weapons by addressing common (mis)conceptions. The analysis discusses 1. AWS as clandestine endeavor that triggers curiosity, 2. AWS triggering both fascination and horror, 3. AWS as rhetorical devices of geopolitical aspirations, 4. Devious interpretations of autonomy and AI, and 5. AWS challenging the relationship between humans and machines.

The article leads thematically to the final, fifth, publication of the dissertation that further engages with the addressed tension between fact and fiction. After the empirical case- studies and conceptual explorations the dissertation revisit the theoretical premises. Can the potential AI futures that are evoked with the case studies be adequately analyzed and countered with future directed heuristics from STS, TA and RRI? I argue that a deeper understanding of hype as an opportunist mode of anticipatory future capture is necessary to capture phenomena that connect AI with notions of overpromising and attention seeking. Surprisingly, little theoretical work exists on the concept of hype. It is mostly colloquially used in everyday language but has not entered or truly been established in research domains. Together with Dr. Max Roßmann from Maastricht university and Dr. Frédérique Bordignon form the École des Ponts ParisTech, I wrote the article *Technology hypes: Practices, approaches and assessments*, which appeared in late 2023 in the *Journal for Technology Assessment in Theory and Practice.* In the article we conceptualize hype, in contrast to vision or expectation as both descriptive and action-guiding. We present different approaches how to empirically study hype as inappropriate exaggeration and the opportunist seeking of attention. In detail, we focus on rhetoric and discourse for the emotional appeal of overpromising language, the playing with temporality and attention spans, or the theatrical Impression management for the creation of followership and collaboration. All these pathways should be understood as preliminary research trajectories of a still scattered research field. To study hype it is revealed

that some national players suggest bold AI future statements and visions and care little about implausibility and epistemological incompleteness – to the contrary, it is exactly this incompleteness that is strategically exploited by some futurists.

THE ARTICLES OF THE DISSERTATION

PART I. DISCURSIVE ARENA CIVIL AI

ARTICLE I. Talking AI into Being: The Narratives and Imaginaries.
of National AI Strategies and their Performative Politics.

ARTICLE II. The 'Trustification' of AI. Forwarding an
Analytical scheme to do justice to a contested phenomenon.

ARTICLE I

Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics[1]

Jascha Bareis & Christian Katzenbach

Abstract

How to integrate artificial intelligence (AI) technologies in the functioning and structures of our society has become a concern of contemporary politics and public debates. In this paper, we investigate national AI strategies as a peculiar form of co-shaping this development, a hybrid of policy and discourse that offers imaginaries, allocates resources, and sets rules. Conceptually, the paper is informed by sociotechnical imaginaries, the sociology of expectations, myths, and the sublime. Empirically we analyze AI policy documents of four key players in the field, namely China, the United States, France, and Germany. The results show that the narrative construction of AI strategies is strikingly similar: they all establish AI as an inevitable and massively disrupting technological development by building on rhetorical devices such as a grand legacy and international competition. Having established this inevitable, yet uncertain, AI future, national leaders proclaim leadership intervention and articulate opportunities and distinct national pathways. While this narrative construction is quite uniform, the respective AI imaginaries are remarkably different, reflecting the vast cultural, political, and economic differences of the countries under study. As governments endow these imaginary pathways with massive resources and investments, they contribute to coproducing the installment of these futures and, thus, yield a performative lock-in function.

Keywords: artificial intelligence, sociotechnical imaginaries, governance, discourse analysis, international comparison

[1] Published 14 July 2021 under CC-BY license in *Science, Technology & Human Values, 47(5), 855-881.* Accessed under: *https://doi.org/10.1177/01622439211030007*. Content and citation style of the original publication have been adopted.

Introduction

Technology is the answer…but what was the question? Cedric Price (1966)

Facing the current rush toward artificial intelligence (AI) by private tech companies such as Google, Facebook, Baidu, or Alibaba, and current public media attention for the subject, governments around the globe have proclaimed to partake in a global AI race (Dutton 2018). In recent years, national AI strategies and regulatory initiatives have been popping up all around the globe. As AI seems to penetrate all spheres of life, governments are on the spot as regulators, articulating potentials, risks, and ethical challenges that go along with current AI developments. Scholars and consultancies have compared and assessed national AI policy papers under the economic frame of "AI competitiveness" and "AI readiness" (Cambrian Futures 2019; Dutton 2018). But these documents do more than merely set rules: they constitute a powerful and peculiar hybrid of policy and discourse. They employ a prose of sober tech-policy, fierce national strategic positioning, and, at the same time, sketch bold visions of public goods and social order enabled through AI.

This paper portrays a comparative qualitative analysis of national AI strategy papers in order to unravel these visions and to deconstruct different idealizations of statehood and algorithmic culture. Notwithstanding the apparent differences in the substantial content of national imaginaries, the key findings suggest a surprising consistency in the narrative of these strategies, converting bold and vague policy talk into a seemingly inevitable technological pathway.

The Integration of AI into Society in Public and Academic Discourse

Topically, this work is situated at the intersection of AI and society that investigates from different angles the coming into being of AI as a key sociotechnical institution of the twenty-first century. Long before the current hype, scholars in sociology and history of science have already studied multiple cycles of hypes and "AI winters" (Bostrom 2014) and extensively

documented and analyzed the social construction of knowledge, scientific practices, and expertise in AI (Woolgar 1985; Courtial and Law 1989; Collins 1993; Suchman 2007). More recent work has stressed that machine learning is far from indifferent to human interaction (Bechmann and Bowker 2019; Castelle 2020), providing detailed ethnographies of technological cultures in AI research (Mackenzie 2017) and mapping the trajectories of competing subfields (Cardon, Cointet, and Mazières 2018). Particularly relevant for the present work, scholars have highlighted the constitutive role of metaphors, myths, and rhetoric: metaphors such as artificial "intelligence" or machine "learning" guide the societal discourse sustainably and fuel fantasies and future visions in the broader public just as much as in expert communities (Campolo and Crawford 2020; Natale and Ballatore 2017). Popular AI discourse also strongly rests on long-standing motifs of human-like machines in mythical storytelling and science fiction (Bory 2019; Cave and Dihal 2019).

In existing studies of media reporting and fictional representation of AI, scholars have identified coverage that primarily showcases the latest high-tech products and services. Here, business actors feature much more often in AI reporting than other stakeholders (Brennen, Howard, and Nielsen 2018; Chuan, Tsai, and Cho 2019; Fast and Horvitz 2017). This industry agenda-setting favors an overhyped vision of AI, resulting in a public focus on potentials of AI and neglecting its actual methodological limitations (Elish and boyd 2018). Recent studies of media coverage of AI in China reveal a similar dominance of the private sector in propagating positive discourses around AI but also stronger government propagation (Zeng, Chan, and Schäfer 2020).

Scholars have also started to track and analyze the recent uptake of regulatory initiatives on AI across the globe but particularly in Europe, Northern America, and Asia (Daly et al. 2019; Niklas and Dencik 2020). This literature analyses regulatory measures and investments, foregrounding ethics as a normative framework (Jobin, Ienca, and Vayena 2019). While this rise of ethical guidelines certainly constitutes a strategic move by the corporate sector to escape actual regulation (Wagner 2018), it also functions as a tool for governance, at least by shaping the very understanding of AI and its normative challenges (Larsson 2020).
In sum, the literature on AI's integration into society articulates a strong role for discourse in shaping the present and future sociotechnical pathways. Interestingly, scholars have not

yet analyzed governmental positioning on AI and its role in future imaginary production. Certainly, governments are impacted by public and private narratives, but, in turn, they are themselves powerful players in shaping our perception and expectation of AI.

Conceptual Frame: Sociotechnical Imaginaries (SIs), Myths, and the Sublime

In this paper, we approach national AI policy and strategy papers as a peculiar hybrid of policy and discourse. They are at the same time tech policy, national strategic positioning, and an imaginary of public and private goods. In most cases, they sketch broad visions and ambitions but also allocate resources to AI research, list already issued policies and regulations, and present roadmaps for future measures and initiatives. Such a complex interplay asks for a conceptual frame that can do justice to this intricate relation of discourse, politics, and technology. For this reason, our research builds on existing concepts in science and technology studies, such as 'SIs,' but also strongly draws on political theory, sociology, anthropology, and communication and rhetoric studies.

In recent years, Science and Technology Studies (STS) has increasingly become interested in the conjunction of discourse and the making of politics and technology (Mager and Katzenbach 2021). Scholars study "expectations and stories about the future" (van Lente and Rip 1998; van Lente 2016), the role of technological innovations, and visionary rhetoric in enterprises (Beckert 2016) and highlight the discursive struggles around "contested futures" (Brown, Rappert, and Webster 2017). Authors have also investigated the role of futurist narratives and myths, particular regarding the internet and online activities (Flichy 2007; Mansell 2012; Mosco 2005). These "vanguard visions" (Hilgartner 2015) and the rhetorics of "pioneer communities" (Hepp 2020) are now receiving increasing attention in studies of the making of digital futures. With even more attention to language and words, scholars in linguistics, media, and communications have looked at metaphors (Lakoff and Johnson 1980) and their relation to technology (Wyatt 2017). In sum, these studies show that novel technology and science discoveries are regularly linked to modernist narratives of progress, especially in liberal capitalist and communist state systems that depend on technology as a means for market innovation and social engineering. In turn, looking at technology narratives

serves as a means to look into desired futures, informing us about societal strivings and aspirations.

At the nexus of politics, discourse, and technology, the concept of SIs (Jasanoff and Kim 2009) has explicitly foregrounded the role of the state. The authors assert that sustaining imaginaries are always "associated with active exercises of state power, such as the selection of development priorities, the allocation of funds, the investment in material infrastructures" (Jasanoff and Kim 2009, 123). While subsequent research has shown that imaginaries are routinely rather multiple, contested, and commodified than uniform visions of the state (Mager and Katzenbach 2021; Jasanoff 2015), the role of the state remains crucial. It has the capacity to structure future expectations by combining powerful measures of issuing regulations and allocating resources with its own narratives and visions. State actors possess the (legitimate) means to sketch future societal pathways and, at the same time, craft influential institutions that define the virtues and vices facilitated by novel technologies and culture.

In the analysis, we substantiate this high-level concept with, firstly, Mosco's (2005) concept of myths as structuration devices for sociotechnical ordering. With Mosco, the power of myths (such as the apparently always imminent advent of "general AI") does not stem from their level of truthfulness: "myths are neither true nor false, but living or dead (…). To understand a myth involves more than proving it to be false. It means figuring out why the myth exists, why it is so important to people, what it means, and what it tells us about people's hopes and dreams" (p. 29). Hence, debunking myths as sole superstition and simple nonsense would disregard their proper social function. Instead, the deconstruction of successful myths brings to the forefront present desires and values as well the underlying power structures. Barthes (1972) pointed out that myths inhabit a concealing and escapist function, serving to bridge contradictions in society and to escape routine everyday life. Most importantly, this implies a process of depoliticization: the narratives of successful myths massively reduce complexity and decouple developments from their social contexts and power structures. In consequence, myths push human and institutional agency to the background by imagining an unconstrained as-if world of possibilities. This rhetorical function, as the analysis will unravel, is very present in SIs of AI.

For reconstructing and explaining the awe that is often evoked by technological progress, Marx (2000) and Nye (2004) have coined the term technological sublime. The Romantics used the figure of the sublime to describe how natural phenomena and the riddles of physics evoke a feeling of overwhelming grandeur and astonishment. During the nineteenth century, with its early engineering masterpieces such as the railway, the sublime is increasingly "directed toward technology or, rather, the technological conquest of matter" (Marx 2000, 197). Evoking this technological sublime embodies the celebration of technological progress and conceals its problems and contradictions (Marx 2000, 207). As the upcoming analysis will show, this figure can be presently found in the historical framing of AI and help to understand how the agency can be shed away from humans and projected onto AI.

Lastly, we will refer to a greater body of literature regarding the sociology of expectations in order to explain the performative role of the articulation of hopes and fears projected on AI (Beckert 2016; van Lente 2016; van Lente and Rip 1998) in the policy texts at hand. When visions around novel S&T projects are announced, they are often embedded in a rhetoric of prospective potentials that innovation sets free. This rhetoric not only enduringly frames the perception of business and customers for a technology but also creates an element of performativity. "Expectations can be seen to be fundamentally 'generative', they guide activities, provide structure and legitimation, attract interest and foster investment" (Borup et al. 2006, 285–286). What begins as a bold promise, as we will see in the rhetoric analysis of the AI imaginaries, can quickly set free a notion of requirement and necessity—a powerful rhetorical motif urging figures to deliver on the promises. In concert, these conceptual frameworks will jointly function as sensitizing concepts for the following analysis that will focus on both the narratives (The Narratives of National AI Strategies: Talking AI into Being section) and the substantial imaginaries (The Imaginaries of National AI Strategies and Their Performative Politics section) articulated in national AI strategy papers.

Methods: Toward an In-depth Discourse Analysis of AI Tech-policy Strategie

In recent years, numerous countries around the world have been advancing national AI strategy papers. In this paper, we focus on the AI strategies of China, the United States, France, and Germany. This choice of countries is not exhaustive (Daly et al. 2019; Niklas and

Dencik 2020), but it entails key players in the field. Their published AI strategies have received broad international attention, they feature industries and companies that are leading in AI tech development, and these countries share a geopolitical and economic positioning in the world that influences AI development far beyond their borders. The United States and China claim leadership in the global AI race; while France and Germany represent the most powerful nation-states and economies in the European Union with distinct approaches to AI deployment.

The strategy documents are special in various regards. Firstly, they are not set in stone but are subject to substantive updates, adjustments, or even radical dismissals and reorientations. Just as in other political fields, tech policy adapts to political situations and is largely affected by changes in government, for example, after the 2016 elections in the United States, where, ever since the Trump delegation took office, a substantially different stance on AI has been taken. Further, AI strategies are often not limited to one condensed official document or even one type of medium alone. Documents that receive the status of a strategy paper can entail summary reports of summit conference proceedings (2018 White House Summit on AI for American Industry [WHSum]; cf. Table 1), announcements of state councils (A Next Generation Artificial Intelligence Development Plan [NgDpl]), or reports by national expert groups (VilRp). These different media and forms of AI strategies already reflect distinct national political institutional cultures and complicate the identification of one single type of document as a reference. Pragmatically, in our analysis, we include any document that was officially labeled and published as an AI strategy document by a current government in charge between 2016 and 2020 in the four countries, needing to fulfill the minimum requirement to contain some policy measures on how to steer AI present and future (an exception is made with the United States which has experienced a very recent power shift with the Biden administration taking over in 2021). A list of the documents we collected and analyzed can be found in Table 1.

Table 1. Overview of Documents Analyzed.

Country	Name of the Document	Date of Publication	Abbreviation
United States	2018 White House Summit on AI for American Industry	May 10, 2018	WHSum
	Summary of the 2018 Department of Defense Artificial Intelligence Strategy	February 12, 2019	DoDAIStr
	Summary of the 2018 National Defense Strategy	January 19, 2018	DoDNDStr
	Executive Order on Maintaining American Leadership in Artificial Intelligence	February 11, 2019	ExOAI
	National Artificial Intelligence Research Resource Task Force	June 10, 2021	AIRRTF
Germany	Nationale KI-Strategie	November 2018	StrKI
	Die Hightech- Strategie 2025	August 2018	HtchStr
	Comments from the Federal Government of the Federal Republic of Germany on the White Paper on Artificial Intelligence—A European Concept for Excellence and Trust	June 29, 2020	GerEUWP
France	The Villani Report	March 8, 2018	VilRp
	Speech Macron at the Collège de France (own translation)	March 20, 2018	SpMcr
	AI for Humanity web page	March 2018	AIfHwebp
China	A Next Generation Artificial Intelligence Development Plan	July 20, 2017	NgDpl
	Three-year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018-2020)	December 14, 2017	3yApl
	White Paper on Artificial Intelligence Standardization	January 24, 2018	WpAI

Methodologically, we place this work in the hermeneutical tradition of the study of technological imaginations (e.g., Verschraegen et al. 2017) and vision assessment (e.g., Grin and Grunwald 2000), stemming from the technology assessment and the larger STS community (see an overview by Konrad et al. 2016). The content-based analysis of rhetorical motives represents an analytical explorative method, building on a rich pool of empirical examples that investigate the narratives, constellations, and process dynamics in the construction of contested futures (e.g., Lösch, Armin, and Meister 2019; Roßmann 2020). As a research design, we employ a cross-national comparison of countries (Jasanoff 2015, 24). Such a comparative approach not only discloses the formation of the articulated narratives and SIs but especially sheds light on the similarities, differences, and particularities found in each national articulation. We employ an interpretative discourse analysis that does not primarily focus on content (policy, funding, or regulation announcements, etc.) but instead focuses on the underlying argumentative meta-structure and the resulting imaginaries. To comprehend this construction process, we take into account rhetorical devices and narrative figures such as the technological sublime, myths, and the performative force of expectations as introduced before.

We display and analyze how policy documents merge a highly interpretative flexible technology cluster such as AI and a rather vague and contested discourse into a seemingly inevitable and sometimes even desirable technological pathway. For this aim, we initially undertook a close reading of all the policy documents listed above, independent of national origin, identifying core issues and themes in the depiction of the current national situation

of AI present and future. Secondly, we clustered these themes, unraveling them as central rhetorical building blocks (the inevitability of AI, the necessity of AI, uncertainty, and leadership), which are present across all countries independent of the resulting national imaginaries. Thirdly, we investigated the relationship among these building blocks, understanding them as a coherent (but not necessarily linear) narrative that leads to the specific AI imaginary of each nation.

The Narratives of National AI Strategies: Talking AI into Being

In this section and the following, we will firstly portray the common narrative building blocks (Between Rupture and Legacy: The Inevitability of AI, International Competitiveness and the Interdependence of Technology and Societal Good: The Necessity of AI, and Uncertainty and Leadership: Articulating Hopes and Fears of Technological Advancement subsections) resulting from our analysis. Thereafter, we briefly sketch the different national imaginaries as projections of political culture and social order enabled through AI (AI for Humanity and a Cybernetic Control System: Different Imaginaries subsection) and their performative effect resulting in potential lock-in pathways (Lock-in, Path-dependency, and Performative Politics subsection; cf. Figure 1).

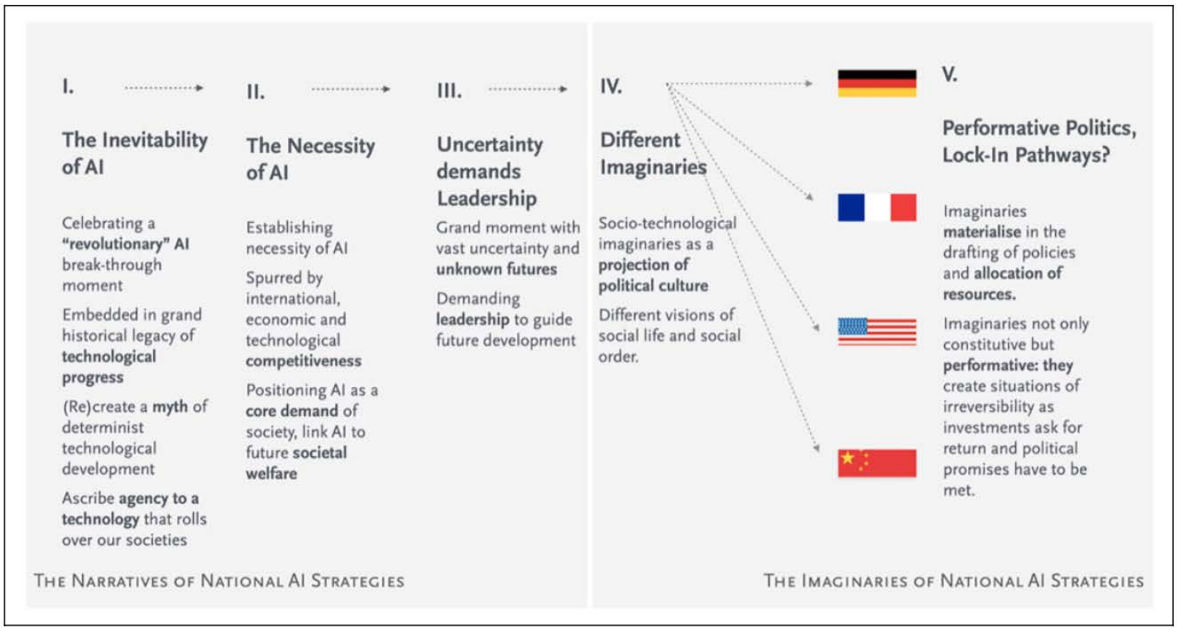


Figure 1. The narratives and imaginaries of national artificial intelligence strategies.

Between Rupture and Legacy: The Inevitability of AI

As a first step of the narrative construction of the AI imaginary, multiple themes can be detected in the strategy papers that convert AI into an inevitable technological pathway.
To set the stage, political leaders situate their societies in a historical context in relation to AI technology. Either such historical context is portrayed as a seemingly unprecedented rupture that transcends any former societal experience or as a rupture that stands in a legacy of past historical transformations. Both historical motives turn current technological AI development into an autonomous agent, a determinist force that breaks over our societies. For example, the Chinese document comments: “The rapid development of artificial intelligence (AI) will profoundly change human society and life and change the world” (NgDpl, 2). Further, AI is portrayed as marking a turning point in world history with US president Trump proclaiming: “We’re on the verge of new technological revolutions that could improve virtually every aspect of our live, create vast new wealth for American workers and families, and open up bold, new frontiers in science, medicine, and communication” (WHSum, quote Trump, 5). Here, AI is depicted as a breakthrough, a revolution, almost a sublime force that lets society enter a new epoch in history. Current transformation is celebrated as a rupture that knows no precedent. In such a context of invoked technological hype, “disjunctive aspects of technological change are often emphasized and continuities with the past are erased from promissory memory” (Borup et al. 2006, 290). Through negating historical continuities, the strategy documents are able to create a myth of a radical break. They suggest a momentum and Zeitgeist of exception, evoking the perception that current transformations will seemingly make everything different, an unforeseen revolution that penetrates every pore of society and makes past reassurances shaky and obsolete. Such denial of history provokes the use of metaphors and images of grandeur that need to underline the current state of exception. Brown et al. (2017) comment in this context: “when the future can no longer be expected to follow on neatly from the past, then imaginative means must be employed” (p. 8). Obscuring past pathways in technological development necessarily purifies (excessively glorifies) and simplifies (reduces or denies social complexity) technological reality. Here, Mosco (2005) stresses: “The denial of history is central to understanding myth as depoliticized speech because to deny history is to remove from

discussion active human agency, the constraints of social structure, and the real world of politics" (p. 35).

Legacy of historical transformations

But the rhetoric of a transcendence of history alone cannot evoke a "breakthrough" perception of AI technology. Analogies and referral to a grand historical legacy equally function to celebrate an upcoming revolution that disrupts humanity. In such a manner, US Deputy Assistant for technological development Kratsios envisions: "Generation after generation, American innovation has benefited our people and the entire world. American oil fueled world industries. American medicine conquered diseases. […] Today, with so many of the mysteries of quantum computing, autonomous systems, and machine learning yet to be discovered, we can take hold of the future and make it our own" (WHSum, 11). And in a similar tone, the Chinese paper states: "AI has become the core driving force for a new round of industrial transformation, [which] will advance the release of the huge energy stored from the previous scientific and technological revolution and industrial transformation, and create a new powerful engine, reconstructing production, distribution, exchange, consumption, and so on (NgDpl, 2 f.).

Here, AI is situated in the linear and coherent promise of historical progress, building upon a legacy of a glorious past. In this context, Jasanoff (2015) comments "technological systems serve on this view a doubly deictic function, pointing back at past cultural achievements and ahead to promising and attainable futures, or to futures to be shunned and avoided" (p. 22). Connecting technological innovations with rhetoric of past revolutions is a strategic move to foster technological celebration, the technological sublime (Marx 2000; Nye 1996). The case of AI sublimation involves hyperbolic statements of technological success, alignment with a national memory of past achievements and a rhetoric of progress that includes the domination over nature or competitors, as well as the conquest over the impossible: "Reference to history and culture can also take the form of analogies to technological success in other fields, which is seen as proof that developments believed to be impossible can actually be realized" (Beckert 2016, 181). At the same time, such accentuation of a historical legacy suggests a notion of human passivity and impotence as we stand still in awe to

contemplate the pathway of a “natural” and “meant to be” historical technological progress that sweeps over our societies.

Such narratives lend agency to technology that transcends human control, confronting society with a seemingly all-pervasive and inevitable development (Brown et al. 2016; Winner 1978) while obscuring the contingencies and power relations of human interaction in the social, political, and economic realm on which any technological development depends. Once an agency is attributed to a technology, and political officials, economic players, and media coverage adapt such discourse, human agency is suddenly reduced to adaption, reaction, or mitigation: “the force implied in this attribution of agency is that one can either ride the wave of advancement or drown in the waves of progress!” (Brown, Rappert, and Webster 2016, 9). French president Macron employs this motive powerfully by stating: “This revolution will not happen in 50 or 60 years, it is happening today, it is really on its track, (…) we have to choose, we have to make certain decisions, given the fact that the technical and the social side is radical and the economic as well” (Speech Macron at the Collège de France [SpMcr]). Nye (2004) highlights that “the most successful of these little narratives are those that present an innovation as not just desirable, but inevitable” (p. 160). Hence, the myth of an inevitable pathway toward AI is created through a play with history that glorifies a seemingly present technological rupture or points at a continuation of a grand legacy, while at the same time negating the role of human agency in such technological development.

International Competitiveness and the Interdependence of Technology and Societal Good: The Necessity of A

he notion of inevitability is fostered not only through the motive of technological determinism, but equally through the pressure of international competitiveness, harnessed within a discourse of capitalist and geopolitical striving for strategic advantage. In the rhetorical construction of an inevitable technological pathway, political leaders establish an interdependent connection between technology advancement, economic performance, and the resilience capabilities of a society. This creates a powerful rhetorical triangle that sheds pivotal attention and necessity to AI, lifting it into a sublime aura of a savior. The Chinese

NgDpl proclaims: “AI has become a new focus of international competition. AI is a strategic technology that will lead in the future; the world’s major developed countries are taking the development of AI as a major strategy to enhance national competitiveness and protect national security” (p. 2). Facing such fierce international competition, the United States and France emphasize their current strategic position in the market. The French Villani report stresses that “It is vital to take advantage of our economy’s comparative advantages and its areas of excellence in order to bolster the French and European artificial intelligence ecosystem” (VilRp, 9). The United States, defending its role as a worldwide leader, makes clear: “America has been the global leader in AI, and the Trump Administration will ensure our great Nation remains the global leader in AI” (WHSum, 8). And further: “Failure to adopt AI will result in legacy systems irrelevant to the defense of our people, eroding cohesion among allies and partners, reduced access to markets that will contribute to a decline in our prosperity and standard of living, and growing challenges to societies that have been built upon individual freedoms” (DoDAIStr, 5). The recent Biden administration, which took over power only this year, continues this narrative by stating: “America’s economic prosperity hinges on foundational investments in our technological leadership” (National Artificial Intelligence Research Resource Task Force [AIRRTF]).

Last but not least, German Hightech strategy paper alerts in a tone of prey and predator: “Even more than in all previous transformations, in this phase of digitalisation the fast beat the slow. The winners will be those who open up new markets early and quickly set their own standards” (Hightechstr, 8 f.).

No matter if packed in a rhetoric of “catching up,” “defending the pole position,” or scenarios of “brute survival,” capitalist competition about market shares and military strivings for geopolitical hegemony fostered through advancement in AI technology are portrayed as of pivotal importance. When such advancement is linked to societal resilience as a whole, technology becomes the crucial tool to master societal challenges or even acts as a yardstick to indicate present status of civilization. Now, technology receives the status of a sublime redeemer that has to be fostered and harnessed. If successful, such a positioning of technology results in an “an aura of indelible pragmatic necessity,” as Winner (1978) notes, and “to ignore these demands, or to leave them insufficiently fulfilled, is to attack the very

foundations on which modern social order rests" (p. 259). Consequently, these narratives elevate AI to become a core demand of society in its entirety, an essential societal good nobody can be deprived of.

Technological advancement acts as an essential pillar of civilizing progress in modern capitalist societies. If a "breakthrough" technology such as AI is detected, while at the same time nations locate themselves in an arena of fierce international competition, politicians magnify the potential of AI to leapfrog economic growth in order to defend (or attain) the nation's global position. Once more, just as with the motive of technological determinism, the advancement of AI now seems vital as the resilience of an entire society depends on it. If the economy, security, and, accordingly, societal order as a whole are at stake, so the narrative suggests, only advancement in AI technology can assure that the current level of living can be maintained and future prosperity secured.

Uncertainty and Leadership: Articulating Hopes and Fears of Technological Advancement

Standing at the verge of such a dramatic historical moment, the consequences are hard to foresee. In the next building block of the construction of AI narratives, national leaders detect prospective potentials, opportunities, challenges, and risks that go along the "inevitable" pathway toward AI and establish a need for leadership.

For China, AI contains the promise of a remedy, projecting hopes of a "technological fix" to social problems: "AI brings new opportunities for social construction. China is currently in the decisive stage of comprehensively constructing a moderately prosperous society. The challenges of population aging, environmental constraints, etc. remain serious" (NgDpl, 3). In consequence, the Chinese government purports the need for strong leadership: "We must strengthen organizational leadership, complete mechanisms, take aim at objectives, keep tasks closely in view, realistically grasp implementation with a spirit of hammering nails, and carry out the blueprint to the end" (NgDpl, 27). Similarly, in the United States, Kratsios sketches a glorious possible future: "Artificial intelligence holds the promise of great benefits for American workers, with the potential to improve safety, increase productivity, and create new industries we can't yet imagine" (WHSum, Speech Kratsios, 9). Here, leadership is more

distributed: “To realize the full potential of AI for the American people, it will require the combined efforts of industry, academia, and government. That is why we are all here today” (WHSum, Speech Kratsios, 8). The German strategy aims to turn the challenges of the transformative rupture of AI into fruitful potentials: “the challenges faced by Germany, as in other countries, involve shaping the structural changes driven by digitalisation and taking place in business, the labour market and society and leveraging the potential which rests in AI technologies” (kiStr, 10). The French strategy stresses the ambivalent character of this AI revolution. President Macron positions himself ready for delivering on these challenges: “(A)s you have understood, you can count on me—I say it here without any innocence—to build the true renaissance that Europe needs” (SpMcr).

While the first two rhetorical themes have downplayed human agency, this third motif brings a new spin to the shared narrative. All strategy papers suggest that future trajectories are undetermined, voicing lofty articulations of hopes and fears rather than clear-cut answers of what the future of AI will bring. This nebulosity serves as a rhetoric that prompts national leaders back into the arena of action. van Lente (2016) highlights that such “statements about future technological performance […] [serve to] mobilize attention, guide efforts and legitimate actions” (p. 46). Upon closer inspection, this spin toward leadership and human agency constitutes a somewhat inconsistent departure from the previous narrative elements of technological determinism and inevitability. If one depicts technological progress as a determinist and historical force by employing vocabulary that suggest human paralysis such as “overwhelming revolution” or “sudden breakthrough,” it is hard to see where there is leeway for decision makers’ agency to shape current and future transformations. Rhetorically, though, the articulation of expectations, hopes, and fears provokes a mobilizing momentum. It serves to open a window of incertitude, which invites for clarification and enables leadership intervention. It offers a suitable opportunity for national leaders to demand initiative and uncritical commitment to coproduce the very futures they envision. Here, “expectations are wishful enactments of a desired future. By performing such futures, they are made real and in this sense expectations can be understood as performative” (Borup et al. 2006, 286).

No matter if a sketched vision or a proclaimed expectation will ever be achieved, it powerfully shapes the discourse. If such political framing is negative, emphasizing the risks and fears that go along the "unstoppable" technological train of progress, then national leaders are put into an intervening role as saviors who can responsibly interfere or at least mitigate worst-case scenarios. Through such rhetoric, also rather less favorable decisions are easy to justify, as confronted with a bleak doomsday scenario (e.g., AI eradicating billions of jobs, AI technology provoking an international arms race), stakeholders are rather willing to bite the bullet. Likewise, though, the myth of a shiny AI future (e.g., the great vision of unprecedented economic growth, the automation of all tedious labor through AI) is a handy means to trigger an uncontested rushing toward a simplified and innocent golden future, often setting aside the social, political, and economic complexities, contradictions, and pitfalls that go along the new innovation.

In sum, AI's political rhetoric about hopes and fears is far from being informative alone. First and foremost, it is constitutive as it frames discourses and (im)possibilities; it is enabling as it allows political activity (also in the face of a looming threat); it can be disguising as it leaves unpleasant societal side-effects and questions about power structures unmentioned and finally also (de)legitimizing, bestowing legitimacy upon political leaders or social institutions—or authorizing certain standpoints or disapproving or condemning others (e.g., cherish technological progress against a "cynical cultural pessimism" or "reactionary Luddism").

The Imaginaries of National AI Strategies and Their Performative Politics

As we have shown, the narrative construction of the national AI strategies are strikingly similar. Yet, their substantial imaginaries are remarkably different, which is probably not surprising given the vast cultural, political, and economic differences of the countries under study. States offer future pathways and at the same time endow these visions with massive resources and investments. As a result, these imaginaries not only reflect on and offer sociotechnical trajectories but, at the same time, coproduce the installment of these futures and, thus, yield a performative function.

AI for Humanity and a Cybernetic Control System: Different Imaginaries

Germany, for example, focuses on AI applications in the manufacturing industry (also branded as AI made in Germany) and promotes an AI imaginary along ethical lines: "We want to use the potential of AI further to improve security, efficiency and sustainability in particularly important fields of application whilst also promoting social and cultural participation, freedom of action and self-determination for each and every citizen" (Nationale KI-Strategie, 9). Here, the German state commits to rather vague normative goals, nonetheless demanding commitment to the promises AI brings along. AI is connected to demands currently en vogue on political agendas, such as security (facing potential cyber and terrorist attacks), efficiency (facing international economic competition), and sustainability (facing the current threat of pollution and global warming). Even though not explained in detail, such terms are linked to liberal core values such as inclusion, freedom of action and autonomy, resembling the stark reference to the German constitutional framework in the German AI strategy papers.

In a similar vein, the French strategy commits to a humanist ethos, stressing to push AI into sectors that enable human flourishing: "[AI] Industrial policy must focus on the main issues and challenges facing our era, including the early detection of pathologies, P4 medicine, medical deserts and zero-emission urban mobility" (AI for Humanity web page). Further, Macron announces, "basically, we return to a new, very Cartesian stage of this faculty of being master and possessor of nature, and it is in this responsibility that we must always situate our action […]. It is a moral responsibility, it is also the guarantee that our democracies will not succumb in some way to an Orwellian syndrome where technology is no longer an instrument of freedom, but a form of control authority" (SpMcr). In grand style, Macron portrays humanity as being at a turning point. The ostentatious presentation of his humanist vision is underlined by figures of philosophy and mythology (Descartes, Prometheus) and serves to create an imaginary of a moral bastion, offering the promise of technological advancement enabling humanist progress. AI is embedded in a philanthropic imaginary to overcome the pressing threats of humanity. It is blessed with an aura comparable to an undeniable fundamental right, a public good, a remedy that can relieve humanity from the vices of our era with the latest innovative technological achievements. Besides such philanthropic narratives, the Villani report claims that inside these

transformative sectors, France can draw on its "economy's comparative advantages and areas of excellence" (VilRp, 9).

The United States takes a remarkably different stance on AI: "Artificial intelligence holds tremendous potential as a tool to empower the American worker, drive growth in American industry, and improve the lives of the American people. Our free-market approach to scientific discovery harnesses the combined strengths of government, industry, and academia, and uniquely positions us to leverage this technology for the betterment of our great nation" (WHSum, 2). Under the Trump administration, the vision of AI is articulated as an act of patriotism, equalizing the technological advancement of the American nation with the advancement of society as a whole. In this context, the term AI serves to unravel essential core values the Trump delegation regards as pivotal, such as empowerment of the American worker, strengthening local industry, or fostering a deregulating free-market approach. In contrast to the French statist vision, the Trump administration aims at removing barriers to AI Innovation "wherever and whenever we can to let American industry, American thinkers, and American workers reach their greatest potential" (speech Kratsios, WHSum, 11). The current Biden administration follows this nationalist narrative by stressing: "The National AI Research Resource will expand access to the resources and tools that fuel AI research and development, opening opportunities for bright minds from across America to pursue the next breakthroughs in science and technology" (AIRRTF). In the US version, AI embodies the free spirit of American scientific ingenuity, the dedication of hardworking people in the rust belt, the competitive economic strength of a proud nation building on a long tradition of narratives of progress and America's culture of greatness (Marx 2000; Nye 1996).

Lastly, the Chinese AI imaginary points again in a different direction, with the Chinese Communist Party depicting AI as a tool for establishing social order and regulation: "Based on the goal of improving people's living standards and quality, speed up and deepen the applications of AI, increase the level of intelligentization of the whole society to form an all-encompassing and ubiquitous intelligent environment" (NgDpl, 18). Further, "AI technologies can accurately sense, forecast, and provide early warning of major situations for infrastructure facilities and social security operations; grasp group cognition and

psychological changes in a timely manner; and take the initiative in decision-making and reactions—which will significantly elevate the capability and level of social governance, playing an irreplaceable role in effectively maintaining social stability" (NgDpl, 3). In order to meet such aims, the Chinese government targets the "smartification" and "intelligentization" of all possible societal fields. In the Chinese strategy papers, AI is interoven with other high-end technological buzzwords such as "smart city," "intelligent robotics," "Industry 4.0," or "facial biometric identification," sketching a totality of AI. Such visions of "data behaviorism" (Rouvroy 2013) or cybernetic governmentality through "environmental-behavioral control" (Krivý 2018) embody a SI where social order is established through a perpetual mode of citizen (self-)monitoring, adaptation, and optimization. The Chinese vision of AI enabling the "construction of public safety and intelligent monitoring and early warning and control system" (NgDpl, 20) echoes Jasanoff's portrayal of a sociotechnical aspiration for "simplification and standardization of human subjects so as to govern them more efficiently" (Jasanoff & Kim 2009, 122).

Lock-in, Path-dependency, and Performative Politics

With their national AI strategies, governments combine the narrative establishment of a particular moment in time that demands leadership (The Narratives of National AI Strategies: Talking AI into Being section) with steering toward particular, country-dependent pathways (AI for Humanity and a Cybernetic Control System: Different Imaginaries subsection). Hence, national leaders seek to convert a field of lofty rhetoric, contingencies, and insecurities into a concrete path of action, aiming at the implementation of their policies through the performance of responsible intervention and leadership. By allocating substantial funding for AI research and business development, establishing normative principles and hard regulation, they constitute the crucial hinge where ideas, announcements, and visions start to materialize in projects, infrastructures, and organizations. Thus, the national AI strategies mark the departure point for country-specific trajectories, driving a process of closure for the integration of AI into society. This creates a process of path dependency that might even lead to lock-in effects down the road.

Borup et al. (2006) write that “after a time, or even rather quickly, expectations may be seen to exhibit certain material and social path dependencies (lock-in or irreversibility)” (p. 293). On the one hand, such a lock-in phenomenon can be understood as a strategic and desirable outcome for political advocates of a technology endeavor, as it embodies a successful manifestation of political will. When implementation has started and path dependencies are taking place, this also means that doubts and fears have been refuted, political critiques and opponents silenced, and political action that pushes into the desired technological direction prevails. Certainly, it is crucial to stress that, notwithstanding the powerful stakeholders that try to forward a SI, such as in the case of AI, their final realization and wide societal embedding will still meet resistance and skepticism, and will meet unforeseen obstacles, ranging from tedious patent litigations to sudden governmental downfalls. Hence, the process of political implementation and social and cultural embedding is anything but a linear progression from tech talk to technological reality, but a myriad of contested interactions.

Nonetheless, once governments proclaim bold promises, they are on the spot to deliver and perform their capabilities. Hence, on the flip side of the path-dependency phenomenon lays the pressure not to disappoint industry and citizens alike. “When expectations are shared they create a pattern into which the actors themselves may be locked” (van Lente and Rip 1998, 217). Such looming risk of lock-in can create additional pressure for the people in charge to deliver on substantial success. Certainly, at this point, national leaders are playing with the point of a costly return. “What starts as an option can be labelled a technical promise, and may subsequently function as a requirement to be achieved, and a necessity for technologists to work on, and for others to support” (van Lente and Rip 1998, 216). Politicians are able to reinforce established and desirable pathways by demanding the commitment of society as a whole to an appealing imaginary, but simultaneously, their reputation is at stake if they fail to reach their proclaimed visions.

Conclusion

How to integrate AI technologies into the functioning and structures of our society has become a concern of contemporary politics and public debates. In this paper, we have

addressed national AI strategies as a peculiar form of co-shaping this development. Constituting a hybrid of policy and discourse, governments offer in these documents broad visions and allocate resources and rules that seek to realize these very visions. We have situated this analysis in the context of approaches relating communication and future technology development such as SIs, the sociology of expectations, myths, and the technological sublime. In the empirical part, we were able to show that the narrative construction of the national AI strategies is strikingly similar: they all establish AI as a given and massively disrupting technical development that will change society and politics fundamentally. In consequence, the necessity to adopt AI across all key sectors of society is portrayed as taken for granted and inevitable. Yet, governments claim agency to shape those developments toward their respective goals along diverse normative principles. While the narrative construction thus is quite uniform, the respective imaginaries that articulate how to integrate AI into society and how to shape future developments are remarkably different. They reflect the vast cultural, political, and economic differences of the countries under study. Since governments offer future pathways in these strategy papers and endow these visions with massive resources and investments, they contribute to coproducing the installment of these futures and, thus, yield a performative function.

By identifying national AI strategies where ideas, announcements, and visions start to materialize in projects, infrastructures, and organizations, we contribute both empirically and conceptually to a better understanding of the nexus of politics, tech development, and discourse. With AI becoming ever more deeply integrated into our societies, we need to closely observe and comment on this process. Recent technological advancements in AI are severely hyped, and governments contribute to this hype, instead of acting as critical watchdogs, soberly assessing the risks and potentials. Their framing of discourses, opinions, and actions are as much enabling as they are restricting, disclosing a double performative, political role. As Powles (2018) comments, “The endgame is always to ‘fix’ A.I. systems, never to use a different system or no system at all. In accepting the existing narratives about A.I., vast zones of contest and imagination are relinquished.” This is the paradox of AI imaginaries: AI tales sound fantastic and trigger our fantasies, though simultaneously they actually undermine political imagination and political practice by raising expectations of a comforting technological fix to structural societal problems. While much of these debates is still quite

controversial, we do seem to witness already a process of closure for a set of fundamental questions—and the national AI strategies certainly contribute to this. Today, AI is established as a key sociotechnical institution; it is considered as taken for granted and inevitable across many sectors already.

With this paper, we set out to systematically analyze the hype production of an emergent technology like AI. Most probably, the analytical scheme at hand is not limited to national AI production alone but can also help to demystify other technological hypes in the past, present, and future such as nanotechnology, quantum computing, and bioengineering. Such transferability is by now of course no more than a further research suggestion that has to be verified—which goes certainly beyond the scope of this paper. While the underlying technological functioning of these technologies is obviously remarkably different, the hopes and fears that are tied to them may be very similar. Future research clearly needs to further reconstruct how AI and other emergent technologies have come into being in the twenty-first century. But this is also the time that we as social science scholars need to contribute to shaping the debate and the actual developments of the specific future forms of technology, because discourse clearly matters.

References

Barthes, Roland. 1972. Mythologies. New York: Farrarm Straus and Giroux.

Bechmann, Anja, and Geoffrey C. Bowker. 2019, February. "Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media." Big Data & Society 6 (1). doi: 10.1177/2053951718819569.

Beckert, Jens. 2016. Imagined Futures. Cambridge, MA: Harvard University Press.

Borup, Mads, Nik Brown, Kornelia Konrad, and Harro van Lente. 2006. "The Sociology of Expectations in Science and Technology." Technology Analysis & Strategic Management 18 (3–4): 285-98. doi: 10.1080/09537320600777002.

Bory, Paolo. 219. "Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo." Convergence 25 (4): 627-42. doi: 10.1177/1354856519829679.

Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press.

Brennen, J. Scott, Philip N. Howard, and Rasmus Kleis Nielsen. 2018. An Industry-led Debate: How UK Media Cover Artificial Intelligence. Oxford, UK: Reuters Institute for the Study of Journalism.

Brown, Nik, Brian Rappert, and Andrewa Webster. 2016. "Introducing Contested Futures: From Looking into the Future to Looking at the Future." In Contested Futures. A Sociology of Prospective Techno-science, edited by Nik Brown, Brian Rappert, and Andrew Webster, 3-20. New York: Routledge.

Brown, Nik, Brian Rappert, and Andrew Webster. 2017. "Introducing Contested Futures: From Looking into the Future to Looking at the Future." In Contested Futures: A Sociology of Prospective Techno-Science, edited by Nik Brown, Brian Rappert, and Andrew Webster, 3-20. New York: Routledge.

Cambrian Futures. 2019. "Nation AI Readiness." Cambrian Group. Accessed June 28, 2021. https://www.cambrian.ai/nair-index.

Campolo, Alexander, and Kate Crawford. 2020. "Enchanted Determinism: Power without Responsibility in Artificial Intelligence." Engaging Science, Technology, and Society 6 (0): 1-19. doi: 10.17351/ests2020.277.

Cardon, Dominique, Jean-Philippe Cointet, and Antoine Mazières. 2018. "Neurons Spike Back: The Invention of Inductive Machines and the Artificial Intelligence Controversy." Réseaux 211 (5): 173. doi: 10.3917/res.211.0173.

Castelle, Michael. 2020. "The Social Lives of Generative Adversarial Networks. "FAT* "20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3373156

Cave, Stephen, and Kanta Dihal. 2019. "Hopes and Fears for Intelligent Machines in Fiction and Reality." Nature Machine Intelligence 1 (2): 74. doi: 10.1038/s42256-019-0020-9

Chuan, Ching-Hua, Wan-Hsiu Sunny Tsai, and Su Yeon Cho. 2019. "Framing Artificial Intelligence in American Newspapers." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society—AIES '19, 339-44 doi: 10.1145/3306618.3314285.

Collins, Harry. 1993. Artificial Experts. Social Knowledge and Intelligent Machines. Cambridge, MA: MIT Press. Accessed June 28, 2021. https://mitpress.mit.edu/books/artificial-experts.

Courtial, Jean-Pierre, and John Law. 1989. "A Co-word Study of Artificial Intelligence." Social

Studies of Science 19 (2): 301-11.
Daly, Angela, Thilo Hagendorff, Hui Li, Monique Mann, Vidushi Marda, Ben Wagner, Wayne
Dutton, Tim. 2018, June 28. “An Overview of National AI Strategies.” Politics AI (Blog). Accessed January 15, 2021. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.
Elish, M. C., and danah boyd. 2018. “Situating Methods in the Magic of Big Data and AI.” Communication Monographs 85 (1): 57-80. doi: 10.1080/03637751.2017.1375130.
Fast, Ethan, and Eric Horvitz. 2017. “Long-term Trends in the Public Perception of Artificial Intelligence.” Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence (AAAI-17). 963-969. Accessed June 28, 2021. https://dl.acm.org/doi/10.5555/3298239.3298381.
Flichy, Patrice. 2007. The Internet Imaginaire. Cambridge, MA: MIT Press.
Grin, John, and Armin Grunwald. 2000. Vision Assessment: Shaping Technology in 21st Century Society: Towards a Repertoire for Technology Assessment. Berlin, Germany: Springer Berlin Heidelberg.
Hepp, Andreas. 2020. “The Fragility of Curating a Pioneer Community: Deep Mediatization and the Spread of the Quantified Self and Maker Movements. International Journal of Cultural Studies 23 (6): 932-50. doi: 10.1177/1367877920922867.
Hilgartner, Stephen. 2015. “Capturing the Imaginary: Vanguards, Visions and the Synthetic Biology Revolution Stephen Hilgartner.” In Science and Democracy, edited by Stephen Hilgartner, Clark Miller, and Rob Hagendijk, 33-55. London, UK: Routledge. doi: 10.4324/9780203564370-7.
Jasanoff, Sheila. 2015. “Future Imperfect: Science, Technology, and the Imaginations of Modernity.” In Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power, edited by Sheila Jasanoff and Sang-Hyun Kim, 1-33. Chicago, IL: University of Chicago Press.
Jasanoff, Sheila, and Sang-Hyun Kim. 2009. “Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea.” Minerva 47 (2): 119. doi: 10.1007/s11024-009-9124-4.
Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The Global Landscape of AI Ethics Guidelines.” Nature Machine Intelligence 1 (9): 389-99. doi: 10.1038/s42256-019-0088-2.
Konrad, Kornelia, Harro van Lente, Chris Groves, and Synthia Selin. 2016. “Performing and Governing the Future in Science and Technology.” In The Handbook of Science and Technology Studies, edited by Ulrike Felt, Rayvon Fouche,Clark A. Miller and Laurel Smith-Doerr, 465-493. Cambridge, MA: MIT Press.
Krivy´, Maroš. 2018. “Towards a Critique of Cybernetic Urbanism: The Smart City and the Society of Control.” Planning Theory 17 (1): 8-30. doi: 10.1177/1473095216645631.
Lakoff, George, and Mark Johnson. 1980. Metaphors We Live By. Chicago, IL: University of Chicago Press. Accessed June 28, 2021. https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html.
Larsson, Stefan. 2020. “On the Governance of Artificial Intelligence through Ethics Guidelines.” Asian Journal of Law and Society 7 (3), 437-451. doi: 10.1017/als.2020.19.
Lösch, Andreas, Grunwald, Armin, and Martin Meister, eds. 2019. Socio-technical Futures Shaping the Present: Empirical Examples and Analytical Challenges. Wiesbaden, Germany: Springer VS.

Mackenzie, Adrian 2017. Machine Learners: Archaeology of a Data Practice. Cambridge, MA: The MIT Press.
Mager, Astrid, and Christian Katzenbach. 2021. “Future Imaginaries in the Making and Governing of Digital Technology: Multiple, Contested, Commodified.” New Media & Society 23 (2): 223-36. doi: 10.1177/1461444820929321.
Mansell, Robin. 2012. Imagining the Internet: Communication, Innovation, and Governance. Oxford, UK: Oxford University Press.
Marx, Leo. 2000. The Machine in the Garden: Technology and the Pastoral Ideal in America. New York: Oxford University Press.
Mosco, Vincent. 2005. The Digital Sublime (MIT Press): Myth, Power, and Cyberspace, New ed. Cambridge, MA: MIT Press.
Natale, Simone, and Andrea Ballatore. 2017. “Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence.” Convergence: The International Journal of Research into New Media Technologies 26 (1): 3-18. doi: 10.1177/1354856517715164.
Niklas, Jedrzej, and Lina Dencik. 2020. “European Artificial Intelligence Policy: Mapping the Institutional Landscape.” Working Paper, Data Justice Lab, University of Cardiff, Cardiff, UK.
Nye, David E. 1996. American Technological Sublime. Cambridge, MA: MIT Press.
Nye, David E. 2004. “Technological Prediction: A Promethean Problem.” In Technological Visions: The Hopes and Fears That Shape New Technologies, edited by Marita Sturken, Douglas Thomas, and Sandra J. Ball-Rokeach, 159-76. Philadelphia, PA: Temple University Press.
Powles, Julia. 2018, December 7. “The Seductive Diversion of “Solving” Bias in Artificial Intelligence.” Medium. Accessed June 28, 2021. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificialintelligence-890df5e5ef53.
Roßmann, Maximillian. 2020. “Vision as Make-believe: How Narratives and Models Represent Sociotechnical Futures.” Journal of Responsible Innovation 8 (1): 70-93.
Rouvroy, Antoinette. 2013, June 3. “The End(s) of Critique: Data Behaviourism versus Due Process.” In Privacy, Due Process and the Computational Turn, edited by Mireille Hildebrandt and Katja de Vries 143-165. London: Routledge. doi: 10.4324/9780203427644-16.
Suchman, L. A. 2007. “Feminist STS and the Sciences of the Artificial.” In New Handbook of Science and Technology Studies, edited by Edward Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, 139-64. Cambridge, MA: MIT Press.
van Lente, Harro. 2016. “Forceful Futures: From Promise to Requirement.” In Contested Futures: A Sociology of Prospective Techno-science, edited by NikBrown, Brian Rappert, and Andrew Webster, 43-64. New York: Routledge.
van Lente, Harro, and Arie Rip. 1998. “Expectations in Technological Developments: An Example of Prospective Structures to Be Filled in by Agency.” In Getting New Technologies Together: Studies in Making Sociotechnical Order, edited by Cornelis Disco and Barend van der Meulen, 203-29. Berlin, Germany: Walter de Gruyter.
Verschraegen, Gert, Frédéric Vandermoere, Luc Braeckmans, and Barbara Segaert. 2017. Imagined Futures in Science, Technology and Society. New York: Taylor & Francis.
Wagner, Ben. 2018. “Ethics as an Escape from Regulation: From “Ethics-washing” To Ethics shopping?” In Being Profiled, edited by Emre Bayamlioglu, Irina Baraliuc, Liisa

Janssens, and Mireille Hildebrandt, 84-89 (Cogitas Ergo Sum: 10 Years of Profiling the European Citizen). Amsterdam, the Netherlands: Amsterdam University Press. doi: 10.2307/j.ctvhrd092.18.

Wei Wang, and Saskia Witteborn. 2019. “Artificial Intelligence, Governance and Ethics: Global Perspectives.” SSRN Scholarly Paper ID 3414805, Social Science Research Network, Rochester, NY. doi: 10.2139/ ssrn.3414805.

Winner, Langdon. 1978. Autonomous Technology: Technics-out-of-control as a Theme in Political Thought. Cambridge, MA: MIT Press.

Woolgar, Steve. 1985. “Why Not a Sociology of Machines? The Case of Sociology and Artificial Intelligence.” Sociology 19 (4). doi: 10.1177/0038038585019004005.

Wyatt, Sally. 2017. “Talking About the Future: Metaphors of the Internet.” In Contested Futures: A Sociology of Prospective Techno-science, edited by NikBrown, Brian Rappert, and Andrew Webster, 109-26. London, UK: Routledge. doi: 10.4324/9781315259420.

Zeng, Jing, Chung-hong Chan, and Mike S. Scha¨fer. 2020. “Contested Chinese Dreams of AI? Public Discourse about Artificial Intelligence on WeChat and People’s Daily Online.” Information, Communication & Society 1-22. doi: 10. 1080/1369118X.2020.1776372.

ARTICLE II

The trustification of AI. Disclosing the bridging pillars that tie trust and AI together[1]

Jascha Bareis

Abstract

Trustworthy artificial intelligence (TAI) is trending high on the political agenda. However, what is actually implied when talking about TAI, and why it is so difficult to achieve, remains insufficiently understood by both academic discourse and current AI policy frameworks. This paper offers an analytical scheme with four different dimensions that constitute TAI: a) A user perspective of AI as a quasi-other; b) AI's embedding in a network of actors from programmers to platform gatekeepers; c) The regulatory role of governance in bridging trust insecurities and deciding on AI value trade-offs; and d) The role of narratives and rhetoric in mediating AI and its conflictual governance processes. It is through the analytical scheme that overlooked aspects and missed regulatory demands around TAI are revealed and can be tackled. Conceptually, this work is situated in disciplinary transgression, dictated by the complexity of the phenomenon of TAI. The paper borrows from multiple inspirations such as phenomenology to reveal AI as a quasi-other we (dis-)trust; Science & Technology Studies (STS) to deconstruct AI's social and rhetorical embedding; as well as political science for pinpointing hegemonial conflicts within regulatory bargaining.

Keywords:

Trustworthy AI, technology governance, conflict theory of state, ethics of AI, public interest AI, science and technology studies

[1] Published 21 May 2024 under CC-BY license in *Big Data & Society, online first*.. Accessed under: https://doi.org/10.1177/205395172412494. Content and citation style of the original publication have been adopted.

Introduction

Trustworthy artificial intelligence (TAI) is trending high on the political agenda. The advancement of artificial intelligence (AI) technology has been endowed with massive investments and great hopes by governments around the world to solve pressing problems in our societies. However, past incidents related to AI have provoked attention and outcry in media and led to hesitation to continue down the path of AI enthusiasm unquestioningly. AI can be misused to manipulate political opinion with deep fakes (van Huijstee et al., 2021). The COMPAS recidivism risk assessment tool used in the US judiciary paradigmatically shows how incidents of bias and discrimination in data processing can aggravate racism and inequality in criminal prosecution (Angwin et al., 2016). Or, while crucial infrastructure becomes ever more automated with AI, issues of safety, robustness and network vulnerability arise from failing systems (McMillan and Varga, 2022). These are only indicative examples that show some salient problems with AI systems.

Such publicly discussed incidents pose a great threat to building and maintaining trust in AI systems and in the institutions that provide these systems and protect users. Faced with these individual and systemic impacts of AI on our societies, regulators are on the spot to carefully weigh the potentials and risks and develop effective policy. As a result, nation states have addressed the urgency of developing policies that address users' ethical concerns while harvesting the economic and efficiency benefits of AI in strategy and position papers (Radu, 2021). However, while there is a growing emphasis on the trust dimension in AI governance in these papers, the pairing of trust and AI is far from intuitive. It invokes first and foremost an unorthodox relationship: It marries a widely technically employed term, AI, with a social one, trust. How to bridge this technical to social domain is not so obvious and straightforwardly answered (see section Pairing trust and AI – a conceptual challenge).

Why should policy makers and researchers care about trust in the governance of a multifaceted technology like AI? First, to understand the general value of trust for technology governance, it is helpful to recognise that distrust in particular can be very costly for society (Hardin, 2002; Warren, 1999). In general, trust relationships are characterised by a state of uncertainty and risk (Luhmann, 1988; Misztal, 1996). If users had perfect knowledge and

control over their technological environment, notions of trust in technology would be redundant. People are willing to give up control if they can be sure that their peers will not act against their interests (Coleman, 1986). Put simply, if one trusts and gives up control, one can save and/or redirect resources. Distrust not only does the opposite, it can be lastingly damaging as it ruins reputations and leads to a great loss of social and economic capital (North, 1990). When distrust spreads and becomes endemic, everyone infected loses. AI scandals illustrate this phenomenon. In the worst cases, users feel betrayed, AI applications are rejected, providers are boycotted, money is burned, and governments' regulatory capacity is questioned. But this also implies that distrust is not always negative. Citizens signalling distrust can also represent a healthy watchdog mechanism for checks and balances, for example by flagging misplaced or badly executed AI systems, regulatory capture or empty rhetoric[2].

Second, and very concretely, TAI plays a pivotal role in the current regulatory debate, as it spearheads regulatory frameworks such as the European AI Act (AIA). Unfortunately, the regulatory approach to trust so far has been rather vague and confusing, lacking definitions and a deeper understanding of trust (section Trustworthy AI in the current landscape: From ethical values to regulatory frameworks).

Third, the current ethical and regulatory debate on TAI is very much fixated on a technical understanding of AI (section Opening technical AI to social dimensions of trust) and its debugging of harmful effects, such as providing computational methods in inspecting models and providing interpretability (see discussion by Páez, 2019; Zednik, 2019; and von Eschenbach, 2021), or de-biasing, discussing trade-offs between algorithmic efficiency and different variations of fairness (Kleinberg et al., 2016; Wong, 2020). This technical debate has its merit, but it lost track of the actual social preconditions that tie trust and AI together.

[2] I follow Duenas-Cid and Calzati (2023) who argue that distrust is not the binary counterpart of trust, implying an opposite end of the same continuum. As they argue with regard to data-driven technologies, trust and distrust must be "regarded as independent yet complementary facets" that coexist (6) and "contribute together to their coming into being in different contexts" (14). Given the limited scope of the paper, I will mainly focus on the relationship of trust and AI but I will still prove their point and show how trust and distrust shape each other's realms and dynamics.

Therefore, this paper responds to current governance initiatives and ethical discussions that invoke trust as an important variable in AI regulation. To be clear from the start: The main aim of this paper is not to assess whether AI is trustworthy or not, but to give an account of the dimensions that need to be considered in order to be able to assess it. Hence, first of all, this paper takes a step back and revisits the concepts of trust and AI. Given the contested relationship between the two phenomena, what are the epistemic dimensions that tie trust and AI together? To answer this research question I forward and execute an analytical scheme based on four pillars: a) AI as a quasi-other; b) AI's embedding in a network of actors from programmers to platform gatekeepers; c) the regulatory role of governance in bridging trust insecurities and deciding on AI value trade-offs; and d) the role of narratives and rhetoric in mediating AI its conflictual governance processes. It is through this systematization that overlooked aspects and missed regulatory demands around TAI are revealed and can be addressed (see Concluding remarks).

This work can be understood as a follow-up on comprehensive systematization works on trust in information and communication technologies (ICT), such as in e-commerce (McKnight et al., 2002), in information systems (Söllner et al., 2016), or in broader readings of technology (Botsman, 2017). However, the complexity of the AI phenomenon requires both different analytical and disciplinary approaches than the ones targeting ICT systems. Therefore, this work borrows from multiple academic viewpoints and concepts. Among other, I refer to phenomenology in order to reveal AI as a quasi-other that we (dis)trust; Science & Technology Studies (STS) to deconstruct the social and rhetorical embedding of AI; and political science to identify hegemonic conflicts in regulatory bargaining. This, admittedly, wide approach is less a scholarly preference but owed to the complexity of the AI phenomenon itself. With disciplinary blinkers, one would miss the constitutive bridging pillars that connect trust with AI. In my approach, I adhere to the agenda of critical algorithmic studies, which is “essentially, founded in a disciplinary transgression“ (Seaver, 2017: 2).

Trustworthy AI in the current landscape: From ethical values to regulatory frameworks

Ethical principles

Recently, there has been a rich landscape of TAI work emerging in both academic debate and governance proposals. The publication of ethical guidelines has reached a scale that is hard to keep track of[3]. High-level principles are published by political bodies and by big Tech companies that aim to ensure a socially desirable implementation of AI, linking ethical values to notions of trustworthiness (EU High-Level Expert Group on AI, 2019; European Commission, 2020a; OECD, 2019). The most dominant approach towards TAI is embedded in the field of ethics. Here, trust is operationalised as a resulting phenomenon that emerges from following a checklist of ethical requirements that need to be 'handled' or 'taken care of'. In this strikingly instrumental understanding of trust, ethicists list values, such as transparency, privacy, accountability, fairness or robustness as fundamental requirements. Kaur et al. (2023) and Reinhardt (2022) undertake great efforts in assembling all the literature of TAI that unites behind each of these single ethical values (see also Simion and Kelp, 2023).

The ethical discourse, even when condensed[4], is descriptively rich but at the same time abundant and abstract, lacking clarity and consensus. Lists of axiomatic AI principles from the public and private sector levitate over the contested reality of society. It is implied that ethical values can be analytically 'isolated', thereby failing to point to the ambivalences and tensions arising between the values (Mittelstadt, 2019). Furthermore, the overall difficulty and reservation to operationalize normative principles and rights into quantitative and measurable scores for governance, while isolating them from their social surrounding and context (Hoffmann, 2019), has led some to bluntly conclude that the discourse of AI ethics is essentially "useless" (Munn, 2022). On a rather poetic remark, Reinhardt (2022) observes that the academic field of trust and AI has turned into "an intellectual land of plenty, a mythological or fictional place where everything is available at any time without conflicts" (741). In conclusion, ethical values may give guidance for better understanding the risks associated with AI but little can be deduced from the ethical discourse in better understanding the phenomenon of TAI.

Governing frameworks

[3] See the huge inventory of ethics guidelines by AlgorithmWatch https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/.
[4] See systematic reviews and frameworks on the multiple ethical guidelines: Floridi & Cowls, 2021; Jobin et al., 2019.

This ethical discourse is flanked by the crafting of a global governance regime around AI. So far, this regime consists of an overlapping ensemble of private standards, normative principle-setting, concrete standardization efforts, as well as the creation of new legal frameworks that shall extend or replace existing (inter-)national legislation (Veale et al., 2023). Supranational bodies such as the OECD (2019) recommended some guiding (albeit again vague) principles for TAI which it would like to see promoted and implemented, taken up by the United Nations which published a more detailed interim Report on "Governing AI for Humanity" in late 2023.

Of all global players, the EU has unquestionably been most proactive in coming up with a coercive and unified framework for establishing TAI[5]. The AIA passed the European Parliament in March 2024 and will come into force by 2026 (European Commission, 2024). The EU commission had initiated the negotiation process in 2019 to develop a distinct European approach to "Excellence and Trust in Artificial Intelligence" (European Commission, 2020b). In the same year, the High-Level Expert Group on AI (HLEG) set the normative foundations for EUs understanding of TAI, forwarding some ethical principles derived from the EU fundamental rights framework (High-Level Expert Group, 2019). The 2020 European Commission White Paper, embedded in a public consultation process, similarly stressed: "As digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it. Trustworthiness is also a prerequisite for its uptake" (European Commission, 2020b: 1), following up with a bold proposal of an "ecosystem of trust".

This AI EU ecosystem of trust builds on three pillars (5):

> "1. it should be lawful, complying with all applicable laws and regulations;
>
> 2. it should be ethical, ensuring adherence to ethical principles and values; and

[5] Other countries like China or the US have also forwarded AI regulatory initiatives, like the 2023 Chinese "Interim Measures of the Management of Generative Artificial Intelligence Services" or the 2023 US executive order on "Safe, Secure and Trustworthy AI". Especially in the US executive order trustworthiness is stressed but also stays ill defined. I my analysis I will especially focus on the European regulatory AI framework as to date, it is the one which is most elaborated.

3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm."

The final text of the AIA comprises these pillars combined with the economic argument of establishing a common AI market integration. The AIA clarifies that "[t]he purpose of this Regulation is to improve the functioning of the internal market and promoting the uptake of human centric and trustworthy artificial intelligence (…)" (European Commission, 2024: 93). In its regulatory paradigm, the AIA combines a principle-based framework of rights with a risk regulatory assessment of harms, while simultaneously aiming at an innovative and internationally competitive AI market (Krarup and Horst, 2023). In its final version the AIA proposal prescribes various instruments of risk regulation, organised around four risk categories, where each AI application is categorised before entering the market.

The notion of trust enters the picture with the classification of high-risk AI systems. They are handled through a self- and third-party conformity assessment (AIA, Article 43). Such assessment builds on the 2020 'self-assessment list for Trustworthy Artificial Intelligence' (ALTAI), which can be understood as a technical and ethical check list (European Commission, 2020a). If this self-assessment or third-party assessment will be enforced in a rigorous and effective way remains disputable, given the general contestability and interpretative vagueness of ethical values and the questionable willingness of profit seeking companies to curtail themselves with higher conformity obligations. Here, users will simply have to trust providers and third-parties. Interestingly, trust is rather featured as a European selling point in the AIA than really being defined. "The Act portrays this declaration of conformity with EU standards as a chief marker of "trustworthiness" (Paul, 2023:12). Thus, it is the entire EU conformity system that is branded as trustworthy, without any explanation of what is essentially meant by trust in the context of AI. It is striking that the entire EU regulatory framework lacks a single definition of trust. As a result, the presentation of TAI in EU documents appears slightly circular. In a nutshell: The EU AI regulation is trustworthy because AI is addressed by the EU. The term TAI lacks semantic quality. As will be shown, this is problematic because regulation risks missing core dimensions of trust that are important for the governance of AI.

Pairing trust and AI – a conceptual challenge

Towards a sincere understanding of trustBefore delving into the different dimensions of trust in AI (section Trust dimensions in AI), the following section clarifies what to actually look for. Trust is not an axiomatic ethical value as the current ethical debate on AI might suggest. To refer to the introductory remarks, trust is a phenomenon that emerges in the social interaction of individuals and collectives characterised by risk and uncertainty. Conceptual and analytical debates on trust focus on the different reasons for entering into trust relationships and on the characteristics of the trust-giver, the trust-taker, and their relationship. Here, trust is generally understood as a social attitude, a normative, mostly emotional expectation towards an entity x and its performance (Hardin, 2002; McLeod, 2021). Trustworthiness, in turn, is a quality or characteristic of entity x and its performance that motivates to provide sufficient reason to justify the attitude of trust (Nickel, 2013). The commonly used analytical scheme to analyse trustworthiness is a three-place relationship: "B is trustworthy for person A with regard to the performance of x" (Nickel et al., 2010: 431). Applying this analytical scheme to the technological domain is neither intuitive nor unproblematic. The dominant approach to trustworthy technology relates to the factor of functionality, which is understood as reliability in performance. "Reliability is a characteristic of an item, expressed by the probability that the item will perform its required function under given conditions for a stated time interval" (Nickel et al., 2010: 433). It should be noted that the connotation of reliability is heavily influenced by an engineering and rational choice perspective that links the performance of technology to the risk of failure, for example, the risk of infrastructure collapsing.

However, many scholars argue that reducing trust to the notion of reliability does not do justice to the true nature of trust, raising the question of whether one should use the concept of trust at all in the context of technology. They link trust to a richer notion that requires some motivation, also known as 'motive-based' theories of trust. These scholars argue that trust must include motives of goodwill and notions of betrayal, thus emphasising emotional involvement (Baier, 1986; Jones, 1996). Others argue that there must be a moral dimension present, such as moral integrity or a person bound by a moral obligation, in order to speak of trust relationships (McLeod, 2002; Nickel, 2007). These broader conceptions of trust defend

trust as an inherently interpersonal phenomenon. Trust is conceptualised as a uniquely human feature, capable of emotions, agency and moral intentions, rather than a phenomenon between objects or technology. The enthusiasm of some thinkers commenting on the pairing of trust and technology is rather reserved. Jones writes: "Trusting is not an attitude that we can adopt toward machinery. I can rely on my computer not to destroy important documents or on my old car to get me from A to B, but my old car is reliable rather than trustworthy. One can only trust things that have wills (…)" (Jones, 1996: 14; see also Ryan 2020 on AI). These reservations about simply transferring interpersonal trust to human-machine trust are instructive for the TAI debate. If one wants to pair trust and AI, one needs to look for features that characterise human-machine relationships beyond reliability.

Opening technical AI to social dimensions of trust

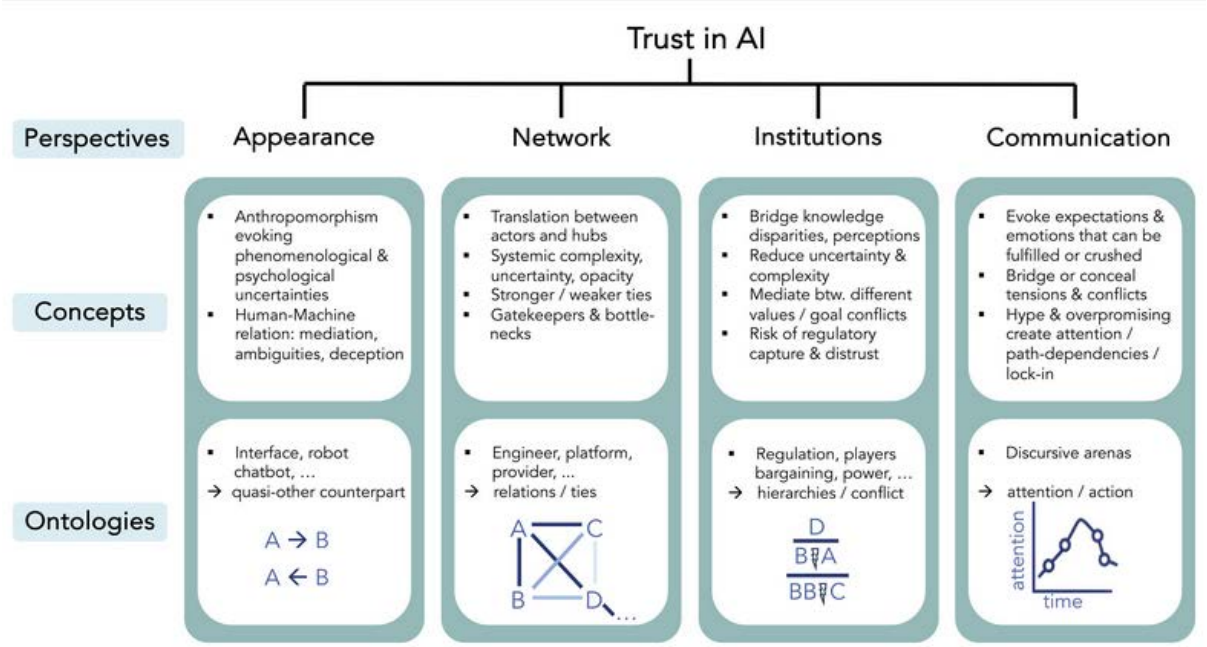
Finding these social and uncertainty realms acknowledges a broader understanding of AI. There is a plethora of definitions of AI coming from academia, corporations, tech gurus and policy papers. Certain features of AI are favoured in certain disciplines, reflecting the diversity of existing AI applications and research. This abundance of discourse has unfortunately led to much confusion around the term in both policy (Folberth et al., 2022) and in public discourse (Natale and Ballatore, 2020) (see also Promoting trustworthy AI through narratives: mediating meaning & attention).

From a technical perspective, AI applications aim to perform some ideal action or reasoning associated with mimicking human tasks and thinking (Krafft et al., 2020). Due to recent technical developments in data processing capabilities and the implementation of statistical learning theory, machine learning (ML) has become the state of the art in AI applications, alongside logic and knowledge-based approaches (Russell and Norvig, 2022). ML relies on great access to data to make robust predictions and to correct performance errors in iterative computational sequences. The technical focus of AI is also dominant in policy papers. For example, the AIA, Art. 3, defines AI as a "machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that (…) can influence physical or virtual environments".

Surprisingly, social environments are not part of the AIA AI definition and that is problematic if one wants to understand the role of trust in the picture. While technical definitions may suggest delimitation and clarity, they fall short of a larger notion when it comes to encompassing AI's relationship to trust. They fail to capture the distinct phenomena that AI applications produce, which arise not so much from algorithmic performativity but the meaning that is ascribed to it. I argue that AI is not only embedded in the social - but is constituted by it. The way AI is perceived and approached by users, embraced by institutions, praised by tech-gurus, and talked about in media points to a constant and complex dynamic between the actual technological developments and the potentials, fears and futures that are associated with it. It is exactly this constant tension between fact and fiction, hype and reality, scandal and breakthrough which is rendering AI so performative as a social phenomenon. I follow a reading that builds on an understanding of AI as situated and relational (Suchman, 2023; Suchman and Weber, 2016; Mackenzie, 2015), reworked and understood by different users and enmeshed in constellations of power. AI is hardly perceived and approached as a clearly articulated, delimited, and external 'thing', 'model' or 'tool' like some technical definitions suggest. Also, in their daily interaction users actually never see code, databases or backends of AI applications. Rather than approaching AI as a self-standing entity that can be generalised ('AI is x'), in this reading AI is woven and negotiated in the everyday realities of users and society, with its applications mediating human relationships, producing intimacies, social orders and knowledge authorities. It is exactly in this dynamic sphere that I will place the analysis of the following sections, as it is here that one can locate the constitutive bridging pillars that tie trust and AI together. The upcoming scheme (see Table 1) should be understood as an offer to policymakers and researchers when they invoke trust relationships with AI, doing justice to the complexity and fragility of the phenomenon. Building trust is challenging, but also rewarding. As outlined in the introduction, respecting the role that (dis)trust plays in the acceptance and rejection of technology is central to designing successful policies.

Trust dimensions in AI

Table 1. Four different trust dimensions that constitute TAI. Visualizing the metastructure of the upcoming analysis



Phenomenological appearance: trusting AI as a quasi-other

From its very beginnings - the foundation of modern AI in the 1950s - AI has been associated with the phenomenon of anthropomorphism: the attribution of human characteristics to objects, behaviours or features - in this case, machines (Salles et al., 2020). In 1966, the computer scientist Joseph Weizenbaum fed his chatbot ELIZA with the DOCTOR script, imitating a Rogerian psychotherapist. ELIZA was a very rudimentary chatbot, programmed to simply rephrase patients' answers as backfeed questions (Güzeldere and Franchi, 1995). Weizenbaum was struck when he observed that his chatbot elicited very emotional and intimate responses from his probands. What has since become known as the 'ELIZA effect' is a powerful demonstration of how humans can project emotions onto machines. The experiment shows that it is not so much the human-like capabilities of algorthmic decision making programs that trigger anthropomorphism (since ELIZA was a very simple software), but their combination with the vast field of human imagination. It is this combination of suggestive human characteristics of a machine with the power of human imagination that

enables the emotional attachment to AI, whether it be social robots, assistive interfaces, or recent large-language-model chat bots like ‘Chat-GPT’ or ‘Gemini’.

Recent academic discourses such as postphenomenology or robot-ethics have elaborated new epistemologies for technological mediation. They develop new concepts of human-machine interaction (Latour, 1994) and technology embodiment (Ihde, 2009; Suchman, 2007); or discuss whether robots appearing in our social world should be understood as moral agents with rights (Loh, 2019; Wallach and Allen, 2008). Without entering into the discussion of whether it being legitimate or helpful to call AI systems moral agents with wills, it is an empirical fact that they increasingly appear human and interact with us as “quasi-others” (Coeckelbergh, 2012: 75). The recent use of AI in the field of personal assistance technologies based on natural language processing, such as Apple's ‘Siri’ or Amazon's ‘Alexa’ (Silva de Barcelos et al., 2020), social robotics applied in the fields of care, elderly and sex services (Scheutz and Arnold, 2016; Sheridan, 2020), or the use of user interfaces at work (Bader and Kaiser, 2019) are very indicative in this regard.

The phenomenological perspective makes clear that AI systems, even if they only simulate human characteristics such as motivations, morals and emotions, can raise expectations of trust. When people interact intimately with AI systems, they embark on fragile social bonds and expose themselves to emotional attachments. In doing so, they are confronted with a core characteristic of trust: the loss of control. When I show intimate emotions, I expose myself vulnerable as I develop expectations that can trigger feelings of validation, resentment or even betrayal. For the motive-based theorists of trust mentioned above, this phenomenological perspective may be frustrating because it refers only to projections and simulations of social beings, but this does not make it any less attractive to many human interactants. Undoubtedly, societies are only at the beginning of an increasing conflation of the real and the virtual, as AI applications are implemented in all kinds of social spheres.

AI as a quasi-other appears not only in social robotics or interfaces, but also with synthetically generated content. The flooding of the internet with deep fakes or factually false content generated by large-language-models has become a major concern in politics. Here, the blurring is deliberate and systematically aimed at disinformation and manipulation of users and the public, threatening the free formation of opinion and the personal integrity of

individuals (Chesney and Citron, 2019; van Huijstee et al., 2021). The weaponisation of suspicion and distrust has already sparked a deliberate military coup in Gabun in January 2019, where a (quite rudimentary) deep fake video of Gabun's President Ali Bongo appearing numb and motionless went viral amid public speculation about his health condition (Washington Post, 2020).

Conclusively, this section stresses that AI as an intersubjective, quasi-other is a pivotal analytical dimension for understanding the relationship between trust and AI. In the face of AI challenging and blurring reality, regulators are on the spot to intervene. So far, the EU AIA imposes transparency duties on the producer of synthetic content, requiring it to be labelled (Art. 52 III). Synthetically produced content will soon increase in quantity and quality and producers will be harder to identify or deliberately remain anonymous villains. Who will be responsible for identifying and proving what is fake or real in the digital world - and will it even be technically possible to distinguish between these states in the future? What content can users trust or must distrust? Current regulatory frameworks fail to address this gap. While the EU's Digital Service Act (DSA) (European Commission, 2022b) prescribes a “notice and take down action” procedure for digital platforms (Art. 14, 14 III, 19), it comes with a caveat. Platforms are not obliged to actively monitor any content and are exempt from liability for the distribution of illegal content as long as they are not aware of it. They wait to be notified by users to flag illegal and offensive content. This, of course, externalises corporate accountability and leaves considerable room for loopholes.

What current TAI governance discussion is missing completely, though, is a reflection of where to draw the line on the role(s) AI should take as quasi-others in very intimate spheres of society such as care, child education or sexuality. It is here where trust relationships are most fragile and people are most exposed and vulnerable. Individuals are already revealing their most intimate selves to AI applications and to much more rudimentary algorithmic systems (see ELIZA). The intrusion of AI into intimate spheres radically puts society's emotional and moral worldviews up for negotiation, as humans are lured out of their comfortable and taken for granted anthropocentric comfort zones. Which boundaries between humans and AI are still legitimate and to be trusted, which even need to be maintained? So far, policymakers have

provided little guidance on these questions, and societies are navigating rather blindly into an increasingly blurring of the analogue and the digital, the authentic and the fake.

Trust the network. AI's social embedding and platformization

The relationship between AI and trust is not only demarcated by an intersubjective and apparent quasi-other. Many factors in a muted and hidden structural background play a key role in trust, embedding an AI application in a network of relationships between different actors. Among others: company leaders, designers, engineers, clickworkers, policy makers, users, and non-users. This extends the network of trust beyond the technological application. Von Eschenbach (2021) notes: “Trust with respect to technology (…) can only be understood in reference to the system as a whole, and each agent's trustworthiness will be judged relative to the differences in roles, interests, and expertise” (1619). The EU HLEG also stresses the importance of a systemic trust account: “Trustworthy AI (…) concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle” (2019: 5). In effect, the notion of trust is extended from AI as an application to a web of different actors involved in the chain of building and delivering a trustworthy AI system.

In addition to the concealed social and technical background processes inherent to the respective AI system, AI applications are embedded in different use contexts and domains. Today, societies are beginning to implement AI in all fields, whether it is work, health, entertainment, military or administration. AI systems act as sorting systems that decide who to hire or not (Laurim et al., 2021), mediate users’ access to information through recommender systems on platforms (Gorwa et al., 2020), and increasingly decide who to kill and who to let live in combat warfare (Abraham, 2024; Asaro, 2012). It is crucial to emphasise that AI systems are not just a technology one uses, but are themselves a governance tool in public policy to establish, manage and enforce social orders. This pervasive form of government by algorithm, which Danaher (2016) coins ‘algocracy’, or Rouvroy and Berns (2013) refer to as ‘algorithmic governmentality’, shows a trend towards AI supporting or even replacing police, military, legislative and administrative action. Another trend in the

embedding of AI is the dominance of social media platforms and marketplaces. There is a growing centralisation around commercial platforms that act as powerful providers, gatekeepers and bottlenecks for AI applications and services. Commercial platforms use AI technology to evaluate, sort and recommend information flows and users. In doing so, platforms pervasively reshape communication relationships and behaviour (Gillespie, 2010; Nitzberg and Zysman, 2022; Srnicek 2017). Through this central position, platforms reconfigure human-AI situatedness (Suchman, 2007), enforcing new modes of interaction, values, spatial and temporal experiences (e.g., intimacy, ubiquity, acceleration). In terms of trust, the use of AI in society, governance and platforms represents an important embedding that needs to be accounted for conceptually. With AI taking on key tasks in the operation and management of platforms, platforms themselves are also theorised as trust mediators (Bodó, 2021b). These virtual meeting places become sites of trust production by matching buyers and sellers, potential sex partners or bridging transactional uncertainties between customers. Undoubtedly, trust can be built here by platforms moderated by AI - but in turn, as Bodo (ibid.) puts it, it is crucial "to inquire whether we can trust technology to produce trust" (2680).

As shown in this section, trust in AI extends from the obvious and apparent AI application to a network of actors and ties. Moreover, it must also be understood as a governance tool for managing social orders, playing a central role in public policy and in the platformisation of widely used digital services. But: How can users control whether this network of relationships embedding AI is trustworthy? They cannot see or understand all the consequences of the specific technical and political choices made by all actors in the design of AI systems. Nor do they have the skills, let alone the information, to grasp whether AI systems are functioning properly and are integer (for example, by not producing biased results or spreading misinformation). In essence, policymakers must consider that users are being presented with an AI end product that remains completely closed and opaque in its design process, its operating mechanisms, and its underlying normative choices.

It seems intuitive that the much-hailed ethical principles of transparency and autonomy are an essential pillar of a TAI standard, at least to counter this myriad of complexity and opacity. However, much recent empirical research shows that evidence is complicated and not as intuitive as ethical guidelines might suggest (Felzmann et al., 2019). In a German study, König

et al. (2022) show that in interaction with personal AI assistants users “do value explainable AI, i.e., high transparency of the AI assistants, [while] this feature barely offsets even a monthly price of 1.99 Euros as compared to no costs” (8). Moreover, Waldman and Martin (2022) show that AI transparency alone does not suffice to judge public policy decisions based on algorithms as legitimate, “countering arguments for greater transparency as a governance solution” (12). They suggest that a human in the decision-making loop is crucial for sensitive areas like policing or judiciary where it is perceived that human capacities and skills crucially matter, which is also supported by Lee (2018). But then again to the contrary, Krügel et al. (2022) show that human oversight does not counter user overconfidence in corrupted algorithms, transforming humans in the loop without digital literacy into “zombies in the loop” (1). While scholarship needs to further explore which arrangements of transparency and human oversight matter in AI contexts, it is already clear that it is not enough to disclose all the different actors and factors that make up the web of trust around an AI application. Realistically, policy makers need to consider that users cannot monitor this myriad and assess the trustworthiness of all actors. To provide TAI, it is essential that users can rely on institutional governance frameworks that establish, maintain and guarantee a trustworthy web of actors. Regulators and their governance role are central to bridging uncertainties. It is within their mandate and competence to implement a regulatory framework that creates systems of trust assurance.

Trust the AI regulatory framework.Governance ensnared between AI interest mediation and value trade-offs

The sociological and institutional literature on trust recognises for long that trust relationships rely on higher-order arrangements that bridge contexts of social uncertainty and knowledge gaps (Misztal, 1996; North, 1990; Sztompka, 1998; Zucker 1985). The complexity of managing different actors influencing TAI demonstrates both the importance and the challenge for public administrations dealing with AI. To date, AI governance modalities make use of both principle-based top-down regulation and market-based self-regulation, using a variety of cooperation and competition logics to govern AI. While the global AI governance landscape is still scattered and evolving, recently, the formation of more coercive regulatory regimes, most

notably at the EU level with the AIA, DSA and Digital Market Act (DMA)[6] (European Commission, 2022a) come into being.

Before delving into policy details, it is important to take a step back and adopt the perspective of public administrations trying to establish trustworthiness for their AI regulatory frameworks and bridge the uncertainty faced by AI users. Their challenge is to manage and balance the different imperatives present in society. These include industry interests for a deregulatory capitalist agenda, the administrations' own internal security and geopolitical interests, while addressing users' concerns about AI and its alignment with existing legal norms and constitutional frameworks. All these imperatives follow different logics and engage with different narratives in the process of AI regulation, making it difficult to co-construct a common understanding of AI, let alone a consensus for appropriate policymaking (König et al., 2021). Recent special issues on the governance of AI (see Büthe et al., 2022; Taeihagh, 2021) have attempted to structure a still young field and aim to find a common language. Here I follow Büthe et al. (2022) that "laws, regulations and other measures to govern AI (…) do not so much reflect inherent characteristics or objective truths about the technology, but reflect political actors' perceptions given those actors' predisposition" (1722).

Instead of talking about different actors in the policy process, however, it is more appropriate to conceptualise the AI policy process as a bargaining field of conflicting players trying to maximise their stakes. This shift in perspective helps to understand the phenomenon of trust and distrust in AI arising from governance frameworks. It is manifested in decisions about value trade-offs that seem inevitable in AI regulation. Politics is caught in a mediating tension, as it has to accommodate different narratives and imperatives of interests that contradict each other in the policy process. The motif of an ensnared state facing a regulatory dilemma has long been propagated by conflict state theorists such as Offe (1972) or Alford and Friedland (1985), and is also present in the hegemony theory of Laclau (1996) and Mouffe (2013). Recent scholarship has aimed to reintroduce agonistic paradigms into technopolitics, mostly in opposition to a perceived dominant deliberative reading of politics in technology assessment (see discussion by Delvenne and Parotte, 2019; Schröder 2019). From an agonistic political perspective, administrations are pressured to consider different narratives and

[6] See footnote 4.

political interests - without taking sides - in order to be perceived as integer, legitimate and trustworthy. Favouring one societal imperative concerning AI (allowing ubiquitous access to user data to support the rise of AI startups) may neglect the concerns of another player (users' concerns about privacy and data autonomy) and undermine the trustworthiness of the administration. In this context, Sztompka (1998) paradoxically speaks of the need for an "institutionalized distrust" (1). After all, it is not surprising that conflicting opinions and interests clash around AI. On the positive side, it can also be read as a constitutive and vital element of democratic political culture. As Bodó (2021a) writes:

> "This competition for the autonomous powers of the state (…) requires the development of complex networks of institutional distrust, which reflect both the distrust among different societal groups with radically divergent and competing interests, and the very real possibility that any of these groups may overtake any of the bodies of the state" (12).

"Overtaking" may have a strong connotation, but issues of regulatory capture, clientelism and outright corruption pose a serious threat to public perceptions of AI regulation and political mandate. This threat is illustrated by the fact that AI regulation faces pervasive value trade-offs. If some stakeholders value a regulatory framework that promotes transparency, corporate accountability, user autonomy & privacy, and progressive fairness standards for vulnerable groups in AI applications, this comes with the caveat of reducing the efficiency and accuracy of those AI applications. For example, designing AI applications to be more explainable (higher interpretability) is time and cost-consuming. It also reduces the complexity of AI systems and curtails their performance output (Baryannis et al., 2019). Or, it has been shown that to make an AI less discriminatory, a programmer must suppress all correlations and proxies associated with a protected category, such as 'gender' or 'age'. This has a significant impact on making an AI model broader and less specific, further being complicated by different fairness principles inherently excluding each other (Kleinberg et al., 2016; Wong, 2020). Higher accuracy means better performance (algorithmic efficiency), but can also lead to disparate impact (more discrimination against vulnerable groups) (Barocas and Selbst, 2016). It goes without saying that higher standards of privacy and corporate accountability would be highly valued by many users, but would be at odds with large data-driven business models of big commercial platforms. Such inevitable trade-offs in AI governance represent an apple of discord, struggling for harder and softer AI regulation, with the risk of producing

inconsistent or partisan regulatory frameworks. The EU's AI regulatory process is a case in point.

Recent reports by the 'Corporate Europe Observatory', 'Transparency International' and 'Euroactive' show how big Tech, corporate think tanks, and trade and business associations are active in blocking and watering down AI regulation in Brussels. Big Tech, largely dominated by US firms, have "spen[t] over € 97 million annually lobbying the EU institutions (…) ahead of pharma, fossil fuels, finance, or chemicals" (Bank et al., 2021: 6). In 2023 industry lobbyists had by far the most meetings with the EU commission on the AIA, featuring 86% of all behind closed-door meetings (73 out of 98 meetings), and were most active in agenda and standard setting (Corporate Europe Observatory, 2023; Kergueno et al., 2021). For the AIA "tech companies have reduced safety obligations, sidelined human rights and anti-discrimination concerns" (Schyns, 2023: 3). Leaked documents strikingly show how companies try to pressure policy makers for a deregulatory agenda by staging narratives like "Big tech is 'irreplaceable' when it comes to problem solving", "we're just defending SMEs and consumers", "Europe wins the tech race against China, or it falls back into the Stone Age" (Bank et al., 2021: 27). In the final round of discussions on the AIA, these lobbying efforts have been directed against the designation of general-purpose AI as a 'high risk' category in the AIA, with industry fearing that it would overburden and stifle innovation with strict conformity assessments. European startups like 'Mistral' and 'Aleph Alpha' teamed up with US big Tech companies and derailed, with direct ties to political executives in France or Germany, the policy-making process on the last meters. Industry managed to water down the binding fundamental right assessment proposed by the European Parliament on general-purpose AI into mere transparency rules (Corporate Europe Observatory, 2023; Hartmann, 2023).

Reports that show such a disproportionate favouring of industry interests can be a blow to public perceptions of AI. If users feel (and truly, a feeling may suffice) that regulation is being framed in such a way that AI regulatory trade-offs favour powerful interests but lack democratic integrity, they may be reluctant to trust it. Problematically, distrust can become diffuse and endemic - and then persistently damaging - when the contacts between policy and an interest group become too close and increasingly indistinguishable. Lobbying and partisan agenda-setting takes place behind the scenes. Unable to identify and address those

responsible, some publics quickly direct their sentiments of distrust towards diffuse upper hierarchies such as 'the system', 'the powerful elites', 'those Eurocrats in Brussels'. The revolving door phenomenon can certainly fuel this perception. This is undoubtedly the case with AI at the EU regulatory level, as "three quarters of all Google and Meta's EU lobbyists have formerly worked for a governmental body at the EU or member state level" (Schyns, 2023: 7). In general, interest trade-offs are not necessarily problematic if regulator communication is transparent and honest. How value trade-offs are communicated and accommodated is an essential feature of managing expectations, hopes and fears around AI. It draws central attention to the discursive dimension of AI, which leads to the final analytical dimension that pairs trust with AI.

Promoting trustworthy AI through narratives: mediating meaning & attention

Trust in AI is strongly mediated by its discursive framing, which creates meaning what to expect from AI, the promises and fears it embodies, and the problems it is supposed to solve. Hence, the societal role which AI shall fulfil is not innate in technical details but is socially constructed and harnessed. Science and technology needs the social narrative to justify itself as valid, legitimate, needed, and strived for. As will be argued, TAI narratives have a dual societal function: they create acceptance, topicality and attract investments around AI, while at the same time silencing and bridging value conflicts and contradictions as assessed in the previous section.

AI is a technology that is very rich from a narrative standpoint. The extensive discursive embedding of AI with human concepts such as 'thinking', 'autonomy' and 'intelligence' shapes perceptions of AI in both public and expert domains. Since its beginning, AI has raised expectations and dreams of exuberant achievements, constantly entertaining the thought of outperforming the human (Campolo and Crawford, 2020; Dandurand et al., 2022; Natale and Ballatore, 2020). These narratives are often embedded in the binary of hopes and fears, or redemption or doom, most concretely embodied in fictional narratives around AI (Cave and Dihal, 2019). But the fictional quickly conflates into the real, with AI myths being echoed in public arenas shaping overall AI sense making (Crépel et al., 2021). Framed perceptions of AI raise expectations that may be frustrated if promises are not kept, negatively influencing perceptions of both the trust-giver (the communicator of promises, such as providers or

regulators) and the then demystified AI systems. The often-exaggerated image that conveys the potential and danger of AI is critical for the realm of trust, as trust relationships are built on emotional expectations. When users are confronted with a discrepancy between exaggerated promises and the actual reality, this can lead to feelings of dishonesty, disappointment and even betrayal.

Given this context, empirical work shows how nation-states and supranational institutions have actively positioned themselves in the AI arena. Administrations portray themselves in an 'AI race' (Cave and ÓhÉigeartaigh, 2018), employing deterministic rhetoric of an 'inevitable' societal path towards AI. This future trajectory is fuelled by rhetorics of TINA (there is no alternative), politically surrendering to the logics of international economic competition. Likewise, societies being constantly shaken by the exhausting reality of crises transforms AI's role from a technology into that of a saviour, nourishing the epic tale of redeeming society from its current structural problems, such as the urban mobility crisis, social inequalities, or climate change. This solutionist aura (Morozov, 2013) that surrounds AI in the political and cultural realm reifies it as given and needed – thereby defining the toolkit to combat socially deeply rooted problems. With the race to AI portrayed as inevitable, a race to AI regulation (Smuha, 2021) is also evoked, pressuring governments to come up with effective regulatory frameworks. However, selling smart AI-based solutions while ignoring deep-rooted social problems can be a pitfall for TAI. The sociology of expectations and STS warn about the risk of such tech-ubiquity leading to path dependencies and lock-ins (Borup et al., 2006; van Lente and Rip, 1998). Managed public expectations of AI can easily turn into demands on governments. As I have argued elsewhere with Bareis and Katzenbach (2022), deconstructing the consistency of national AI imaginaries: "Once governments proclaim bold promises, they are on the spot to deliver and perform their capabilities" (874). The praise of technology talk becomes performative and can increase the pressure not to disappoint users. Stakeholders are playing with the trust of AI-users if raised expectations are not met and promises prove empty – or scandals shatter the before hailed AI solutions.

In general, not only AI but also TAI has become a buzzword in politics. As outlined in the section before, the EU has framed its entire regulatory framework with the emblem of TAI. While TAI remains completely under-defined, it functions as an empty signifier that has its

political function. By deploying the TAI frame, the EU Commission can rhetorically accommodate stakeholders and their conflicting interests and unify a contested field of actors in a seemingly harmonised and consensual regulatory framework. From the outset, "AI industry can read 'trustworthiness' as a call for robustness, while ethicists and legal experts can simultaneously imagine that the document puts forward the agenda of making AI development more ethical and lawful" (Stamboliev and Christiaens, 2024:6). Thus, TAI functions as a unifier to bridge different interests, but this comes with a significant caveat: the carving out of what TAI actually entails. This semantic emptiness may even be cherished and promoted by political actors, but of course it would then lack any substance and meaningful content. Worse, if these empty signifiers are revealed as a strategy to obscure power structures in regulatory processes, the blow to TAI and AI governance bodies can be substantial.

Concluding remarks

Trustworthy AI (TAI) has recently been widely employed in the context of AI regulation and in ethical debates around AI. This paper aims to structure and advance the debate, doing justice to a complex socio-political phenomenon that has suffered from being reduced to a semantically carved-out buzzword. This paper argues that the actual requirements for linking trust and AI are demanding – but also rewarding. Rather than following the dominant path in AI research of linking trust to ethical principles such as fairness, transparency, or privacy, or to technical properties such as robustness, efficiency, or accuracy, I hope to have shown that the phenomenon of TAI (while certainly being influenced by these) mobilises larger epistemic and social dimensions. Any technical approach to de-biasing, auditing, or making AI more transparent has its merits, but ultimately falls short of capturing and doing justice to the variously situated realms that constitute TAI. These include a) AI as an intersubjective relationship, with trust being negotiated through AI as a quasi-other; b) the embedding of AI in a network of actors from programmers to platform gatekeepers; c) the regulatory role of governance in bridging trust uncertainties and deciding on AI value trade-offs; and d) the role of narratives and rhetoric in mediating AI and conflictual AI governance processes (see overview Table 1). Admittedly, the analytical scheme is a heuristic and therefore necessarily abstract. I have executed each dimension with regard to AI in this paper, but in reality, they

easily conflate. Some work more in the foreground with interfaces and materialities, others are enmeshed and implicit in power-relations and hierarchies, or framed by conversations about AI Hollywood blockbusters or technical policy results. However, for policy makers and researchers, the analytical scheme has its merit as it structures a scattered debate, points to regulatory requirements and brings clarity for further research trajectories.

Given the regulatory perspective, first, one must state that there are clear policy gaps in the European regulatory acts (other international proposals are still in the making) like the AIA, DSA and DMS. This concerns a questionable self-assessment and third-party risk assessment approach, or insufficient accountability duties for the identification and labelling of AI-generated synthetic content on platforms and search engines. With synthetically generated content flooding the internet, there is an increasing societal disorientation to what extent the blurring of the authentic and factual with the fake and false is socially and politically acceptable. This especially concerns AI applications in fields where users are most vulnerable such as care, education or sexuality.

Second, recent scholarship around internet regulation theorized governance as an open-end reflexive coordination in a complex network of social ties, “ordering processes from the bottom-up rather than proceeding from regulatory structures” (Hofmann et al., 2017: 1413). This actor-network inspired governance perspective serves well to bring all actors who are involved in AI production and distribution to the foreground, but understates the very nature of power and political bargaining between these actors. Hofmann et al. state that governance, here understood as coordination, “becomes reflexive when ordinary interactions break down or become problematic” (ibid.: 1414). This implied deliberative take of governing a complex network would misconceive the nature of hierarchical politics, though. Rather than leaning on a reflexive notion of politcs, I have put forward an agonistic picture of AI governance, depicting strivings for hegemony and agenda setting between players in deciding upon value trade-offs. This perspective serves to understand the political dimension in installing trust or provoking distrust in AI, tackling issues of regulatory capture or revolving doors. These phenomena, of course, are not only limited to AI but also emerge alongside other regulations. Not surprisingly, though, it is especially prevalent when big Tech is aiming to make big money.

Third, I indicated that the carving out of TAI may not only be the consequence of a scattered debate but also depicts political strategy. I have highlighted the role of discourses and narratives for trust in AI, managing expectations through playing around with hopes and fears. It is revealing that transparency, integrity and honesty have such a low standing in political processes. The fact that the implementation of AI involves value trade-offs is not the fault of policymakers - but the euphemised way in which it is presented, not to mention the unbalanced and hidden lobbying that is allowed to take place, certainly is. Every trade-off with AI has its benefits and perils for society, and these can and should be fully and transparently articulated – and publicly discussed. This would actually relieve politicians of much of the pressure to sugar-coat bad deals and spare them from manoeuvring themselves into rhetorical traps they then struggle to escape. Clarifying the stakes, the actors and their interests is in itself a transparency value that could substantially (re)build trust in political processes and, consequently, in their regulatory objects – in this case, AI.

By disentangling the relationship between trust and AI, this scholarship situates itself within the agenda of critical policy studies (Paul, 2022) and critical algorithm studies (Seaver, 2017). To successfully (dis-)integrate AI for the benefit of all, an understanding of how algorithmic phenomena shape, maintain and challenge society and its order is a pivotal precondition. This understanding calls for disciplinary transgression where needed to disclose how the technical is inscribed, mediated and practised in the social.

References

Abraham Y (2024) 'Lavender': The AI machine directing Israel's bombing spree in Gaza. +97 Magazine. Accessed from: https://www.972mag.com/lavender-ai-israeli-army-gaza/.

Alford RR and Friedland R (1985) Powers of Theory: Capitalism, the State, and Democracy.Cambridge: Cambridge University Press.

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias. Ethics of Data and Analytics, 254–264. Auerbach Publications.

Asaro P (2012) On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. International Review of the Red Cross 94(886): 687–709.

Bader V and Kaiser S (2019) Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. Organization 26(5): 655–672.

Baier A (1986) Trust and antitrust. Ethics 96(2): 231–260.

Bank M, Duffy F, Leyendecker V, et al. (2021) The Lobby Network: Big Tech's Web of Influence in the EU. Brussels: Corporate Europe Observatory. Retrieved from: https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu.

Bareis J and Katzenbach C (2022) Talking AI into being: The nar- ratives and imaginaries of national AI strategies and their performative politics. Science, Technology, & Human Values 47(5): 855–881.

Barocas S and Selbst AD (2016) Big data's disparate impact. California Law Review 104: 671.

Baryannis G, Dani S and Antoniou G (2019) Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Future Generation Computer Systems 101: 993–1004.

Bodó B (2021a) The commodification of trust. Blockchain & Society Policy Research Lab Research Nodes, 1. DOI: 10.2139/ssrn.3843707.

Bodó B (2021b) Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. New Media & Society 23(9): 2668–2690.

Borup M, Brown N, Konrad K, et al. (2006) The sociology of expectations in science and technology. Technology Analysis & Strategic Management 18(3–4): 285–298.

Botsman R (2017) Who Can You Trust? How Technology Brought Us Together–And Why It Could Drive Us Apart. London: Penguin UK.

Büthe T, Djeffal C, Lütge C, et al. (2022) Governing AI – attempt- ing to herd cats? Introduction to the special issue on the governance of artificial intelligence. Journal of European Public Policy 29(11): 1721–1752.

Campolo A and Crawford K (2020) Enchanted determinism: Power without responsibility in artificial intelligence. Engaging Science, Technology, and Society 6(0): 1–19.

Cave S and Dihal K (2019) Hopes and fears for intelligent machines in fiction and reality. Nature Machine Intelligence 1(2): 74.

Cave S and ÓhÉigeartaigh SS (2018) An AI race for strategic advantage: rhetoric and risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 36-40).

Chesney B and Citron D (2019) Deep fakes: A looming challenge for privacy, democracy and national security. California Law Review 107: 1753.

Coeckelbergh M (2012) Can we trust robots? Ethics and Information Technology 14(1): 53–60.

Coleman JS (1986) Social theory, social research, and a theory of action. American Journal of Sociology 91(6): 1309–1335.

Corporate Europe Observatory (2023) Byte by byte. How big Tech undermined the AI Act. Accessed from: https:// corporateeurope.org/en/2023/11/byte-byte.

Crépel M, Do S, Cointet JP, et al. (2021) Mapping AI Issues in Media Through NLP Methods. In CHR (pp. 77-91).

Danaher J (2016) The threat of algocracy: Reality, resistance and accommodation. Philosophy &

Technology 29: 245–268.
Dandurand G, Blottière M, Jorandon G, et al. (2022) Training the News: Coverage of Canada's AI Hype Cycle (2012–2021). INRS - Urbanisation Culture Société.
Delvenne P and Parotte C (2019) Breaking the myth of neutrality: Technology assessment has politics, technology assessment as politics. Technological Forecasting and Social Change 139: 64–72.
Duenas-Cid D and Calzati S (2023) Dis/Trust and data-driven technologies. Internet Policy Review 12(4): 1–23.
EU High-Level Expert Group on AI (2019) Ethics Guidelines for Trustworthy Artificial Intelligence. Accessed from: https:// digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
European Commission (2020a) Assessment List for Trustworthy AI (ALTAI). Accessed from: https://digital-strategy.ec.europa.eu/ en/library/assessment-list-trustworthy-artificial-intelligence-altai- self-assessment.
European Commission (2020b) On Artificial Intelligence – A European approach to excellence and trust. Accessed from: https: /commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
European Commission (2022a) Digital Markets Act (DMA). Accessed from: https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925.
European Commission (2022b) Digital Service Act (DSA). Accessed from: https://eur-lex.europa.eu/legal-content/EN/ TXT/?toc=OJ%3AL%3A2022%3A277%3ATOC&uri=uriserv %3AOJ.L_.2022.277.01.0001.01.ENG.
European Commission (2024) Artificial Intelligence Act (AIA). Accessed from: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206.
Felzmann H, Villaronga EF, Lutz C, et al. (2019) Transparency you can trust: Transparency requirements for artificial intelli- gence between legal norms and contextual concerns. Big Data & Society 6(1). DOI: 2053951719860542.
Floridi L and Cowls J (2021) A unified framework of five princi- ples for AI in society. Harvard Data Science Review 1(1). DOI: 10.1162/99608f92.8cd550d1.
Folberth A, Jahnel J, Bareis J, et al. (2022) Tackling problems, harvesting benefits–A systematic review of the regulatory debate around AI. arXiv preprint arXiv:2209.05468.
Gillespie T (2010) The politics of 'platforms'. New Media & Society 12(3): 347–364.
Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7(1).
Güzeldere G and Franchi S (1995) Dialogues with colorful "per- sonalities" of early AI. Stanford Humanities Review 4(2): 161–169.
Hardin R (2002) Trust and Trustworthiness. New York: Russell Sage Foundation.
Hartmann T (2023) AI Act: French government accused of being influenced by lobbyist with conflict interests. www.euractiv. com. https://www.euractiv.com/section/artificial-intelligence/ news/ai-act-french-government-accused-of-being-influenced- by-lobbyist-with-conflict-of-interests/.
Hoffmann R (2019) Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 22(7): 900–915.
Hofmann J, Katzenbach C and Gollatz K (2017) Between coord- ination and regulation: Finding the governance in internet gov- ernance. New Media & Society 19(9): 1406–1423.
Ihde D (2009) Postphenomenology and Technoscience. The Peking University Lectures. Peking: Peking University Press. Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9):389–399.
Jones K (1996) Trust as an affective attitude. Ethics 107: 4–25.
Kaur D, Uslu S, Rittichier KJ, et al. (2023) Trustworthy artificial intelligence: A review. ACM

Computing Surveys 55(2): 1–38. Kergueno R, Aiossa N, Pearson L, et al. (2021) Deep pockets, open doors: Big tech lobbying in Brussels. Transparency International EU: 1–21.
Kleinberg J, Mullainathan S and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv pre- print arXiv:1609.05807.
König H, Baumann MF and Coenen C (2021) Emerging technolo- gies and innovation—hopes for and obstacles to inclusive soci- etal co-construction. Sustainability 13(23): 13197.
König PD, Wurster S and Siewert MB (2022) Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis. Big Data & Society 9(1): 205395172110696.
Krafft PM, Young M, Katell M, et al. (2020) Defining AI in policy versus practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 72-78.
Krarup T and Horst M (2023) European artificial intelligence policy as digital single market making. Big Data & Society 10(1): 20539517231153811.
Krügel S, Ostermaier A and Uhl M (2022) Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. Philosophy & Technology 35(1): 611.
Laclau E (1996) Emancipation(s). New York: Verso.
Latour B (1994) On technical mediation. Common Knowledge 3(2): 29–64.
Laurim V, Arpaci S, Prommegger B, et al. (2021) Computer, whom should I hire? Acceptance criteria for artificial intelligence in the recruitment process. Proceedings of the 54th Hawaii international conference on system sciences, p. 5495.
Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society 5(1). DOI: 10.1177/2053951718756684.
Loh J (2019) Maschinenethik und Roboterethik. Handbuch Maschinenethik. Heidelberg: Springer, 75–93.
Luhmann N (1988) Familiarity, confidence, trust: Problems and alternatives. In: Gambetta D (eds) Trust: Making and Breaking Cooperative Relations. Oxford: Basil Blackwell, 94–107.
Mackenzie A (2015) The production of prediction: What does machine learning want? European Journal of Cultural Studies 18(4-5): 429–445.
McKnight DH, Choudhury V and Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. Information Systems Research 13(3): 334–359.
McLeod C (2002) Self-Trust and Reproductive Autonomy. Cambridge: MIT Press.
McLeod C (2021) Trust. The Stanford Encyclopedia of Philosophy (Fall 2021 Edition). Available at:https://plato. stanford.edu/archives/fall2021/entries/trust/.
McMillan L and Varga L (2022) A review of the use of artificial intelligence methods in infrastructure systems. Engineering Applications of Artificial Intelligence 116: 105472.
Misztal BA (1996) Trust in Modern Societies: The Search for the Bases of Social Order. Cambridge:Polity Press.
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nature Machine Intelligence 1(11): 501–507.
Morozov E (2013) To Save Everything, Click Here: The Folly of Technological Solutionism. New York: Public Affairs.
Mouffe C (2013) Agonistics: Thinking the World Politically. London: Verso.
Munn L (2022) The uselessness of AI ethics. AI and Ethics, 1-9.
Natale S and Ballatore A (2020) Imagining the thinking machine: Technological myths and the rise of artificial intelligence. Convergence: The International Journal of Research into New Media Technologies 26(1): 3–18.
Nickel PJ (2007) Trust and obligation-ascription. Ethical Theory and Moral Practice 10(3): 309–319.

Nickel PJ (2013) Trust in technological systems. In: de Vries MJ, Hansson SO and Meijers AWM (eds) Norms in Technology. Dordrecht: Springer, 223–237.
Nickel PJ, Franssen M and Kroes P (2010) Can we make sense of the notion of trustworthy technology? Knowledge, Technology & Policy 23(3): 429–444.
Nitzberg M and Zysman J (2022) Algorithms, data, and platforms: The diverse challenges of governing AI. Journal of European Public Policy 29(11): 1753–1778.
North DC (1990) Institutions, Institutional Change and Economic Performance. Cambridge: Cambridge University Press.
OECD (2019) OECD AI Principles - Recommendation of the Council on Artificial Intelligence. Accessed from: https:// oecd.ai/en/ai-principles.
Offe C (1972) Strukturprobleme des kapitalistischen Staates. Frankfurt a. M.: Suhrkamp.
Páez A (2019) The pragmatic turn in explainable artificial intelli- gence (XAI). Minds and Machines: 29(3): 441–459.
Paul R (2022) Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. Critical Policy Studies 16(4): 497–509.
Paul R (2023) European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a com- petitive common AI market. Regulation & Governance. DOI: 10.1111/rego.12563.
Radu R (2021) Steering the governance of artificial intelligence: National strategies in perspective. Policy and Society 40(2): 178–193.
Reinhardt K (2022) Trust and trustworthiness in AI ethics. AI and Ethics, 1-10.
Rouvroy A and Berns T (2013) Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation? Réseaux 177: 163–196.
Russell S and Norvig P (2022) Artificial Intelligence: A Modern Approach, 4th Global ed.
Ryan M (2020) In AI we trust: Ethics, artificial intelligence, and reliability. Science and Engineering Ethics 26(5): 2749–2767.
Salles A, Evers K and Farisco M (2020) Anthropomorphism in AI. AJOB neuroscience 11(2): 88–95.
Scheutz M and Arnold T (2016, March) Are we ready for sex robots? In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 351-358). IEEE.
Schröder JV (2019) Das Politische in der Technikfolgena- bschätzung: Reflexionen mit der pluralen, radikalen Demokratietheorie von Laclau und Mouffe. TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis/Journal for Technology Assessment in Theory and Practice 28(3): 62–67.
Schyns C (2023) The lobbying ghost in the machine. Corporate Europe Observatory. Brussels, Belgium. Accessed from: https: /corporateeurope.org/en/2023/02/lobbying-ghost-machine.
Seaver N (2017) Algorithms as culture: Some tactics for the eth- nography of algorithmic systems. Big Data & Society 4(2). DOI: 10.1177/2053951717738104.
Sheridan TB (2020) A review of recent research in social robotics. Current Opinion in Psychology 36: 7–12. Silva de Barcelos A, Gomes MM, da Costa CA, et al. (2020)Intelligent personal assistants: A systematic literature review. Expert Systems with Applications 147: 113193.
Simion M and Kelp C (2023) Trustworthy artificial intelligence. Asian Journal of Philosophy 2(1): 8.
Smuha NA (2021) From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. Law, Innovation and Technology 13(1): 57–84.
Söllner M, Hoffmann A and Leimeister JM (2016) Why different trust relationships matter for information systems users. European Journal of Information Systems 25(3): 274–287.

Srnicek N (2017) Platform Capitalism. London: Polity. Stamboliev E and Christiaens T (2024) How empty is trustworthy AI? A discourse analysis of the ethics guidelines of trustworthy AI. Critical Policy Studies 1–18. DOI: 10.1080/ 19460171.2024.2315431.
Suchman L (2007) Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge: Cambridge University Press.
Suchman L (2023) The uncontroversial 'thingness' of AI. Big Data & Society, 10(2). DOI: 10.1177/20539517231206794.
Suchman L and Weber J (2016) Human-machine autonomies. In: Bhuta N, Beck S, Geis R, et al. (eds) Autonomous Weapons Systems: Law, Ethics, Policy. Cambridge: Cambridge University Press, 75–102.
Sztompka P (1998) Trust, distrust and two paradoxes of democ- racy. European Journal of Social Theory 1(1): 19–32.
Taeihagh A (2021) Governance of artificial intelligence. Policy and Society 40(2): 137–157.
van Huijstee M, van Boheemen P and Das D (2021) Tackling deepfakes in European policy.
van Lente H and Rip A (1998) Expectations in technological devel- opments: An example of prospective structures to be filled in by agency. In: Disco C and Meulen B (eds) Getting New Technologies Together: Studies in Making Sociotechnical Order. Berlin, Germany: Walter de Gruyter, 203–230.
Veale M, Matus K and Gorwa R (2023) AI and Global Governance: Modalities, Rationales, Tensions. Annual Review of Law and Social Science 19.
von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology 34(4): 1607–1622.
Waldman A and Martin K (2022) Governing algorithmic deci- sions: The role of decision importance and governance on per- ceived legitimacy of algorithmic decisions. Big Data & Society 9(1). DOI: 10.1177/20539517221100449.
Wallach W and Allen C (2008) Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press.
Warren ME (ed.) (1999) Democracy and Trust. Cambridge: Cambridge University Press.
Washington Post (2020, February 13) The suspicious video that helped spark an attempted coup in Gabon | The Fact Checker [Video]. YouTube. https://www.youtube.com/watch?v=F5vzKs4z1dc.
Wong P (2020) Democratizing algorithmic fairness. Philosophy & Technology 33(2): 225–244.
Zednik C (2019) Solving the black box problem: A normative framework for explainable artificial intelligence. Philosophy & Technology 34: 265–288.
Zucker LG (1985) Production of trust: Institutional sources of eco- nomic structure, 1840 to 1920. In: Cummings LL and Staw B (eds) Research in Organizational Behavior. Greenwich, CT: JAI Press, 53–110.

PART II. DISCURSIVE ARENA MILITARY AI

ARTICLE III. "Autonomous Weapons" as a Geopolitical Signifier: US and Chinese Military Strategies as Means of Political Communication.

ARTICLE IV. The realities of autonomous weapons: Hedging a hybrid space of fact and fiction.

ARTICLE III

"Autonomous Weapon" as a Geopolitical Signifier: US and Chinese Military Strategies as Means of Political Communication.[1]

Thomas Christian Bächle & Jascha Bareis

Abstract

"Autonomous weapon systems" (AWS) have been subject to intense discussions for years. Numerous political, academic and legal actors are debating their consequences, with many calling for strict regulation or even a global ban. Surprisingly, it often remains unclear which technologies the term AWS refers to and also in what sense these systems can be characterised as autonomous at all. Despite being feared by many, weapons that are completely self-governing and beyond human control are more of a conceptual possibility than an actual military reality. As will be argued, the conflicting interpretations of AWS are largely the result of the diverse meanings that are constructed in political discourses. These interpretations convert specific understandings of AI into strategic assets and consequently hinder the establishment of common ethical standards and legal regulations. In particular, this article looks at the publicly available military AI strategies and position papers by China and the USA. It analyses how AWS technologies, understood as evoking sociotechnical imaginaries, are politicised to serve particular national interests. The article presents the current theoretical debate, which has sought to find a functional definition of AWS that is sufficiently unambiguous for regulatory or military contexts. Approaching AWS as a phenomenon that is embedded in a particular sociotechnical imaginary, however, flags up the ways in which nation states portray themselves as part of a global AI race, competing over economic, military and geopolitical advantages. Nation states do not just enforce their geopolitical ambitions through a fierce *realpolitik* rhetoric but also play around with ambiguities in definitions. This especially holds true for China and the USA, since they are

[1] Published 02 September 2022 under CC-BY license in *European Journal of Futures Research*, **10**, 20 (2022). Accessed under: https://doi.org/10.1186/s40309-022-00202-w. Content and citation style of the original publication have been adopted.

regarded and regard themselves as hegemonic antagonists, presenting competing self-conceptions that are apparent in their histories, political doctrines and identities. The way they showcase their AI-driven military prowess indicates an ambivalent rhetoric of legal sobriety, tech-regulation and aggressive national dominance. AWS take on the role of signifiers that are employed to foster political legitimacy or to spark deliberate confusion and deterrence.

Keywords:

autonomous weapon systems, China, USA, deterrence, hybrid warfare

Introduction

The development of the so-called autonomous weapon systems (AWS) has been the subject of intense discussions for years. Numerous political, academic and legal institutions and actors are debating the consequences and risks that may arise with these technologies, in particular their ethical, social and political implications, and many have called for strict regulation or even a global ban [1,2,3].

In these public debates, the attribute "lethal" is sometimes added to the term AWS, underlining the potential severity of the consequences this technology entails. Surprisingly, and despite the urgent need to deal with "Lethal Autonomous Weapon Systems" (LAWS), it is often unclear which technologies the term (L)AWS[2] primarily refers to, or even in what sense these systems can be characterised as "autonomous" at all. The associated definitions describe a range of phenomena, from landmines to combat drones, from close-in weapon systems (CIWS) to humanoid robot soldiers or purely virtual cyber weapons. Besides this terminological ambiguity, it is inherently unclear in what sense or to what degree these systems can be characterised as "autonomous" at all. Even though the development of automatic or semi-autonomous capabilities is generally advancing, fully autonomous weapons that are completely beyond human control—which is the reason why they are feared by many—largely represents a conceptual possibility at present rather than an actual military reality ("Technical definitions of autonomy and autonomous weapons systems" section).

While the current debate around the possibility and functionality of AWS is certainly not a novel phenomenon but one that has also been highly influenced by fictional works of the past [4], it has regained prominence in recent decades with technological advancements in artificial intelligence (AI), especially with accelerating machine learning (ML) data processing capabilities. Civil societal initiatives [5, 6], scientists [7, 8] and political bodies[3] have raised political concerns about emerging "intelligent" and "autonomous" weapon systems with

[2] Cf. "The challenges of defining autonomous weapon systems" section for more details on the attribute "lethal"

[3] See e.g. the debate at the CCW discussed below, "AWS as geopolitical signifiers: strategies in political communication in China and the United States of America" section.

lethal capabilities that go beyond human control. As much as the debate has been guided by the agendas of different stakeholders pursuing (de-)regulation, the discourse around AWS has developed alongside other genres such as doomsday stories in journalism, Hollywood cinema or science-fiction literature, which exploit the idea around looming "killer robots". Besides promoting a certain idea of what AWS are and what they are capable of, they also intensify the political debate by adding a high degree of urgency.

As will be argued, the conflicting interpretations of AWS are largely the result of diverse meanings that are constructed in political discourses. They convert a specific understanding of AI into strategic assets and, as a political consequence, hinder the establishment of common international ethical standards and legal regulation. Hence, the perspective we present not only reveals AWS to be powerful signifiers of political culture but also shows how they are instruments employed to foster political legitimacy or to spark deliberate confusion and deterrence between rival states.

In particular, this article looks at the publicly available military AI strategies and position papers by China and the USA and, informed by sociotechnical imaginaries [9, 10], analyses how this technology is politicised to serve particular national roles and interests. The ways these two nations showcase their AI-driven military prowess sends out unmistakable messages about national dominance and a desired geopolitical order. The ways in which nation states portray themselves as part of a global AI race, competing over economic, military, and political advantages, become obvious. This especially holds true for China and the USA, since they are regarded, and regard themselves, not only as international hegemons, but also as antagonists, promoting competing self-conceptions that are apparent in their histories, political doctrines and identities.

In turn, the analytical focus on these hegemonic powers will inform European debates on AWS, since these discussions are far from representing one unified stance. Identifying the similarities and differences between China and the USA makes it possible to recognise prototypical patterns, which at the same time puts the multitude of different AWS positions among European nations into a larger global perspective[4]. The analysis explicitly focuses on

[4] See "Methodology" section and the conclusion for details on the French and German initiatives at the CCW, as they take an important role in the UN discussions on regulating AWS.

military strategy documents in an effort to complete the picture of national AI aspirations and more general public discourses. Specifically, this subdomain of AWS imaginaries was chosen because it brings to the fore the deliberate meanings voiced by military actors in order to utilise them as part of political communication.

The article first dissects the current academic debate regarding a definition of AWS that would be sufficiently unambiguous for regulatory or military contexts; key issues in this debate have been concepts such as "autonomy", "degree of human control" or a "functional understanding of AWS" ("The challenges of defining autonomous weapon systems" section). It is the meaning of these AWS-related concepts that, among other dimensions, constitutes the reference point in the geopolitical arena between the USA and China. They not only provide information about technical details but can be utilised to fulfil specific functions in asserting national interests. In order to be able to approach and analyse AWS from this *realpolitik* perspective, we introduce the concept of the "sociotechnical imaginary" (SI) as the theoretical frame ("Approaching autonomous weapons embedded in sociotechnical imaginaries" section). The "Methods" section follows ("Methodology" section), where we showcase the empirical material, consisting of position papers taken from the debate at the United Nations (UN) Convention on Certain Conventional Weapons (CCW[5]) and standpoint papers published by the executive ministries of both nations. The analysis sections portray AWS as geopolitical signifiers and approach the strategies as a form of political communication that is pursued as part of military AI imaginaries ("Military doctrines, autonomous weapons and AI imaginaries" section). AWS are a central element of the goals both nations pursue in the realm of geopolitical communication. Differing definitions and normative understandings of AWS are deliberately employed to serve national interests and, consequently, make it more difficult to reach a UN regulatory consensus ("Technological definitions and normative understandings of AWS" section).

The challenges of defining autonomous weapon systems

[5] The long version reads are as follows: The Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects.

The different approaches to defining AWS constitute an arena of competing interpretations of what the technology is capable of and, above all, which reference points to consider in order to regulate specific capabilities. While the current debates on autonomous weapon systems mainly focus on regulatory questions, military simulation games or political and tactical scenarios, the power of interpretation over what AWS are and what capabilities they comprise remains contested. These questions neither simply refer to a problem in engineering nor are they of a purely conceptual nature but also borrow from the realm of fiction. It is essential to acknowledge that the prerogative of shaping the meaning of the technology creates both semantic and political dominance—and states take advantage of this opportunity.

In order to narrow down a comprehensible understanding, three different approaches can be roughly distinguished: The first focuses on the attribute of "autonomous", which evokes a wide array of traditional associations with the *concept of autonomy*; the second approach takes into account different *degrees of human control* over the automated processes and in doing so addresses questions of human/machine interaction. While it is obvious that both definitional approaches are directly interwoven—in a complimentary fashion even, since the more autonomous the machines are, the less human control can be exercised—they still refer to distinct conceptual meanings and traditions. The third and most recent strategy promotes a primarily *functional understanding of AWS* that focuses on the actual capabilities and seeks to transcend essentialist definitions that are more concerned with the innate conceptual qualities of the technology.

Technical definitions of autonomy and autonomous weapon systems

One possible way of defining the concept of autonomy is to look at it as a technically determining and distinguishing feature; indeed, this already seems self-evident from the attribute "autonomous" alone. In this sense, an "autonomous" weapon system is one that, "based on conclusions derived from gathered information and preprogrammed constraints, is capable of independently selecting and engaging targets" [11]. While automated systems are only "triggered", in this understanding, such systems can independently "select" and "engage" different targets, based on case-specific information.

The concept of autonomy is widely used in philosophy, psychology, human cognition and other disciplines and carries (often contested and contradictory) meanings that range from anthropocentric understandings to political contexts or aesthetics [12,13,14]. It has become a quite commonplace term in AI discourses, where it commonly evokes clear associations with characteristics such as independence, intelligence, self-governance, the ability to learn and adapt (e.g. orientation in unknown, unstructured and dynamic environments) or the execution of self-determined decisions. Its ubiquitous use, however, which also shapes non-expert debates on AI, has contributed to the erosion of its semantic qualities.

Even when one narrows down the concept to a more specific technical sense, ambiguities persist. Bradshaw et al. emphasise that there are two different understandings of autonomy in the context of machines: "In the first sense, it denotes self-sufficiency—the capability of an entity to take care of itself. The second sense refers to the quality of self-directedness or freedom from outside control. [...] It should be evident that independence from outside control does not entail the self-sufficiency of an autonomous machine. Nor do a machine's autonomous capabilities guarantee that it will be allowed to operate in a self-directed manner. In fact, human-machine systems involve a dynamic balance of self-sufficiency and self-directedness". At the same time, since no entity can be seen as completely independent of its environment, the term autonomous system would in a strict sense even count as a "misnomer" [15].

Furthermore, the different interpretations of machine autonomy in the context of AWS are usually embedded in either optimistic or dystopian discourses, which in turn firmly shape the understandings of autonomy as well, in particular the sense of "what autonomous machines can and cannot do" [16]. It is exactly this interpretative openness that make AWS an important reference point in the politico-strategic interactions of rivalling states, which are continuously struggling for a clear definition. A consensus on what can be regarded as an autonomous weapon is seen as a first step towards the legally binding regulation of these technologies.[6]

[6] As Ekelhof continues to point out: "This linguistic indeterminacy has not withheld States from claiming consensus on a number of fundamental points; in fact, it may even have facilitated the development of these two consensus claims: (1) International law applies to autonomous weapons, and (2) some form of human involvement is necessary to ensure the lawful use of autonomous weapons. This may seem a notable

The discussions on these semantic issues are held at the regular (annual or biannual) meetings that take place between participating state parties on the protocols of the CCW that was adopted in 1980 (cf. “Methodology” section) [17]. Politically, the terminological ambivalence and polysemy opens the door for disagreement at the CCW on how to define “autonomy” (cf. “Technological definitions and normative understandings of AWS” section). This, as a direct consequence, has also led to the failure to regulate autonomous weapons [18]. Paradoxically, even a common terminology can make the discourse on AWS more complicated, “when the terms involved lack consistent interpretations”. The often metaphorical use of “autonomy” and its ambiguities creates uncertainty when military robots are treated as black boxes. Only when understanding human decision-making processes in the design, production and programming of autonomous machines, questions of agency and responsibility can intelligibly be discussed [19].

This is why solely looking for ways to determine AWS in terms of the concept of autonomy cannot be sufficient, as the label “autonomous” evokes a whole spectrum of meanings that nonetheless does not present us with finite categorical distinctions. Even the more precise term of the so-called *technical autonomy* refers to a continuum, a point that becomes obvious by the necessity to employ auxiliary vocabulary such as “*semi*-autonomous”. In short, the term “autonomous” alone—even when defined technologically and hence relatively unequivocally as the “capabilities” of AWS—is not enough to grasp its complexity, since the weapon must necessarily also be understood in the ways it presents itself in manifold contexts.

Definitions focusing on the degree of human control over supposedly autonomous systems

Another approach to defining AWS involves determining the degree of human control over a weapon system that remains unaffected despite a higher degree of automation. In particular, it was the notation of *in*, *on* and *out of the loop*—emphatically used not in the sense of an inherent technical property, but in relation to human agency—that gained prominence in the debate. “In-the-loop” refers to directly executed control by humans (an action must be

achievement, but the linguistic indeterminacies that exist in this context inevitably turn these professed commonalities amongst High Contracting Parties into empty—or at least weakened—claims of consensus. This raises the question [...]: what do these claims actually mean?” [88]

initiated), “on-the-loop” refers to systems whose actions can be prevented or aborted by human intervention, and, finally, “out-of-the-loop” is the term commonly used for systems that no longer require human control but whose processes are, most of the time, nonetheless still monitored by human agents.

According to this approach, weapon systems are to be called autonomous if they reduce the possibility of human intervention to a minimum, up to the point where they no longer require or even allow human control at all. It reflects a *relational* understanding of autonomous weapons in terms of the possibility of human intervention and agency and hence can be seen as part of a broader model conceptualising human/machine relationships.

In practice though, the focus on a relational understanding of agency and automation still comes with terminological challenges. One of these challenges refers to the vague distinction between automation and autonomy. As Sauer notes: “After all, automatic systems, targeting humans at borders or automatically firing back at the source of incoming munitions, already raise questions relevant to the autonomy debate” [20]. Similarly, defining the degree of human control as a continuum is at best a measurement metric, as the complex interactions cannot always be clearly attributed to either the human or the machine [21]. Further complicating this approach, this distinction says little about the “autonomy” of the system itself, but at best classifies the possibilities for curtailing it [11]. In other words, even a weapon system that *could* be called autonomous in a technical sense (cf. “Technical definitions of autonomy and autonomous weapons systems” section) can easily fall short of these expectations and functional properties if it is deliberatively limited and curtailed in a context that is controlled by humans (see “Technological definitions and normative understandings of AWS” section for a detailed analysis of the terminology used in US national strategy papers regarding AWS). The questions remain whether it makes sense to regard it as “autonomous” and even whether the attribute conveys a useful meaning at all. As Ekelhof comments, “any consensus among states, academia, NGOs, and other commentators involved in diplomatic efforts under the auspices of the CCW ... seems to be grounded in the idea that all weapons should be subject to “meaningful human control” (or a similar standard). This intuitively appealing concept immediately gained traction, although at a familiar legal-political cost:

nobody knows what the concept actually means in practice" [22] (see also "Technological definition: United States of America" section).

Functional approaches to what "autonomous weapon systems" can and cannot do

The terminological vagueness partly explains more recent endeavours to find a functional definition of AWS. As we will see, however, these task-specific approaches rearrange and combine the above-discussed conceptual and relational understandings and engender their own problems, even though they are trying to break them down to actual functionalities in practical settings.

The most common way to a functional understanding of autonomous weapons at present is a task-based focus on "selecting" and "engaging" a target, which reframes the above definitions but puts stronger emphasis on what these functions comprise and entail in specific practical settings. The US Department of Defense (DoD) has defined an AWS as a "weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation" [US.PosP1] (see "AWS as geopolitical signifiers: Strategies in political communication in China and the United States of America" section for a detailed analysis). This approach is gaining traction and political acceptance. The International Committee of the Red Cross defines AWS as "any weapon system with autonomy in the critical functions of target selection and target engagement". That is a weapon system that can select (i.e. detect and identify) and attack (i.e. use force against, neutralise, damage or destroy) targets without human intervention [23], with commentators emphasising that the "adoption of the ICRC's definition—or one like it—" was "strongly advisable" paired with a call for "concerted response by the international community" to the continued developments of these kinds of weapons [24].

Ekelhof notes that the "main focus within this definition lies on the so-called critical functions of target *selection* and *attack* and the absence or lack of *human intervention* in relation to the system's *autonomy*" [25]. Both target selection (sometimes meaning the mere distinction between combatants and non-combatants, sometimes referring to larger planning processes)

and attack (raising the questions of what constitutes an individual attack or when exactly it starts and ends), in the end, bear their own ambiguities, albeit in a less obvious manner [26].

Even efforts to define AWS by focusing on specific tasks fail to establish a common ground that would clearly distinguish them from previous weapon systems while at the same time meeting the expectation of unambiguously pinpointing their functionality. Both "autonomy" and "meaningful human control" are volatile signifiers. The same, however, applies to automated tasks that are interpreted as constitutive of autonomous weapons, since these tasks are embedded in military practices, infrastructures and concrete situations that eventually determine the effects and degrees of autonomy. In other words, the contexts produce the conditions under which the agency of an autonomous weapon is determined.[7]

Hopes that a functional, task-oriented definition of AWS (specifically singling out target selection and engagement) would neatly solve the ambiguity problem are bound to be disappointed. Even the more precise terminology is subjected to political discourses, in which different actors deliberately utilise diverging meanings, interpretations and definitions to pursue particular political and geostrategic interests. This picture is complicated even further by voices from outside the political realm, which claim that the current AWS technologies are not sophisticated enough to reasonably draw conclusions regarding their practical, legal or ethical consequences [27].

Both the conceptual and the task-centric approaches lead into a semantic recursion, as in all cases—irrespective of the level of theoretical abstraction—the necessity to agree on a static meaning of the terms cannot be met. One important issue usually neglected in these debates is the challenge of translating these terms back and forth between languages that are situated in vastly differing terminological and conceptual traditions (Bächle TC, Champion SC: Autonomous weapon systems. Journalistic discourses in China, forthcoming)[8]. These cultural differences manifest themselves in larger imaginaries, promoting specific expectations, hopes

[7] This, of course, does not trivialise the questions of human agency (as necessary fail safe) or human responsibility (that must not be delegated to machines).

[8] The term "autonomous/autonomy" and with it the term "autonomous weapon" does not have a direct equivalent in Mandarin (Bächle TC, Champion SC: Autonomous weapon systems. Journalistic discourses in China, forthcoming).

and fears around new technologies. They are promoted by fictional texts but also by public discourses. For AWS, the attribute "lethal" is a case in point here. By the addition of the L in LAWS, the term comes to emphasise that these technologies are in line with expectations associated with the so-called killer robots, evoking specific cultural images. These images foreground the potential harm that is associated with autonomous weapons outside of human control, extending to fears of looming destruction of all humanity. The following section particularly addresses the role of larger sociotechnical imaginaries that shape and determine the ways AWS become meaningful technologies[9].

Approaching autonomous weapons embedded in sociotechnical imaginaries.

Continuously re-semanticising or bluntly denying the mere possibility of a reasonable discourse on AWS and their effects are two ways that are used to drag out the efforts to find effective regulation. At the same time, AWS are only one of the many fields that shape the AI race between state actors and are rhetorically embedded in larger sociotechnical imaginations that are actively politicised. This becomes especially apparent when we look at the two self-proclaimed superpowers, China and the USA, both of which are striving for global dominance. In both instances, the national discourses around AWS act as signifiers that reveal projections of social, cultural and institutional imaginations. Arguably, these discourses not only function as meaningful narratives but also as effective instruments of geopolitical power (e.g. with the intention of deterrence) to enforce specific interests grounded in *realpolitik*.

The contradictory and contested meanings that are associated with and at the same time constitutive of AWS are embedded in larger narrative structures that in this article are regarded as an expression of vivid "sociotechnical imaginaries" [10][10]. In a well-known and influential understanding of "sociotechnical imaginaries", Jasanoff defines them as

[9] Ekelhof recounts that autonomous weapons "were first discussed in the Human Rights Council in 2013 under the name "Lethal Autonomous Robotics" and later that year the topic (referred to as "Lethal Autonomous Weapons Systems") was placed on the United Nations Convention on Certain Conventional Weapons' (CCW) agenda for the year 2014 [89] Despite the meaning that is (probably deliberately) communicated with the use of "lethal" as an attribute, "the military has long applied the word "lethality" to anything that could make weapons more effective, not just the weapons themselves but also to training, methods, intel support systems and more" [90].

[10] For individual analyses of sociotechnical imaginaries see Jasanoff and Kim [10], for case studies regarding the interconnectedness of knowledge production, technologies and social order see e.g. Hilgartner et al. [91].

"collectively held, institutionally stabilised, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology" [28][11]. In the continuation of this definition, the "desired futures" are juxtaposed with the "shared fears of harms that might be incurred through invention and innovation"—imaginings between utopia and dystopia—perfectly align with the discursive positions guiding the debates on AWS.

A vast body of research in the wake of Jasanoff's initial coining of the concept has shown that imaginaries powerfully set boundaries to our futures, "shaping terrains of choices, and thereby actions" [29]. The diversification in approaches and research objects associated with the concept shows that SIs must always be understood as an open, contested and dynamic field influenced by a multitude of discursive arenas and players [10, 29, 30]. For example, AWS imaginaries are often influenced by popular culture, fiction or images used in journalism and inspired by more general assumptions about AI (Bächle TC, Bareis J, Ernst C (eds): The realities of autonomous weapons, forthcoming). The utopian and dystopian frames of reference for AI portray it as a kind of superintelligence with the potential to exceed (human) biology and unleash beneficial effects [31] (e.g. see the Chinese employment of "evolution" in "Technological definitions and normative understandings of AWS" section in the context of AWS), while the rise of technological agency poses grave ethical challenges [32]. AI can be seen as "a key sociotechnical institution of the twenty-first century" with state actors playing a pivotal role in shaping the images in which it is portrayed [33]. AI is strongly associated with specific meanings—and myths—about technological futures [34].

Sociotechnical imaginaries (SIs) mediate between the contested realms of fact and fiction and "allow actors to move beyond inherited thought patterns and categories and into an as if-world different from the present reality" [35]. This also applies to AWS and the foregrounding of science-fiction inspired technologies such as robots, which are promoted on the basis that they will play a vital part in future warfare [36, 37]. Today's "military-entertainment complex" [38] is increasingly blurring the lines between the realities of war and its representation in

[11] The definition continues as follows: "It goes without saying that imaginations of desirable and desired futures correlate, tacitly or explicitly, with the obverse—shared fears of harms that might be incurred through invention and innovation, or of course the failure to innovate. The interplay between positive and negative imaginings—between utopia and dystopia—is a connecting theme throughout this volume " [28].

popular culture (such as war games, which include tactics or threat scenarios). Drones, for example, have become emblematic of a specific type of warfare that has become mediated, remoted, networked, decentred and de-personalised. The particular "aesthetics" of drone images is represented in the arts, literature and film, and in this form, they also enter the public discourse, reifying a particular visual aesthetics of war [39]. This is a continuation of a type of consumable war that is televised, providing live images to the home viewer [40], a type of mediated war whose most recent iterations focus on cyberwars or the "weaponisation of social media" [41].

Paradoxically, it is exactly in this context of uncertainty—in which reality, imagination, possibility and fiction are conflated—that AWS become highly momentous, in particular when political or military decision-making comes to be based on potential or virtual scenarios [42, 43]. The debates around autonomous weapons usually focus on their legal, political or ethical ramifications. The foundation of these works is (at least in part) also based on those potential or virtual scenarios [44]. An ethical problem contributes to constructing, disseminating and maintaining a specific understanding of "(lethal) autonomous weapons" in popular culture, politics, journalism or research [45, 46]. Ethical debates are a major arena for imagining AWS, controversially situated between positions that argue that warfare could even become more "humane" (by more effectively adhering to international law and respecting human rights), when the actual acts of war are left to machines [3, 5] and voices of AI and robotics researchers warning of dire consequences [7].

Approaching AWS as part of the AI imaginations that are deliberately promoted by nation states, it becomes obvious how countries actively portray themselves as part of a global technology race, competing over economic, military and geopolitical advantages. These AWS meanings are part of larger narratives of national identity, interwoven with specific ideologies, ideas of military self-assurance and pride, which in turn are utilised with the communicative goals of deterrence towards political adversaries.

Comparing the USA and China in this regard is particularly fruitful and demonstrative, as they not only locate themselves in the geopolitical arena as rivals with their own interests, but also fundamentally oppose each other in their self-portrayal. This spans from guiding principles in state doctrine, political systems or general canons of values to the origin myths of these

nations, representing competing self-conceptions that are apparent in their diverging histories and political identities.

Schematically, the USA's hunger for greatness, exceptionalism and aspiration to take the role of a global hegemon contrasts with China's confidently proclaimed ideal of a harmonised and stable society. AI is in both cases regarded as a means to realise these socio-political ideals, with supremacy achieved by technological prowess being a shared theme for both. The conceptual ambiguity of autonomous weapon systems makes their representation and interpretations a flexible tool in political communication. AWS can be seen as a proxy for the respective understanding of the world by China and the USA, a form of national self-assurance through technology.

Methodology

In this paper, we focus on the AWS strategies of China and the USA. Obviously, this selection of countries is not exhaustive, but as discussed above, it lends itself to overtly competing, even antagonistic stances of ideological, institutional and historical narratives of the two nations. These differences become particularly apparent in the military guidelines for reaching their respective ambitions. Both China and the USA position themselves as global leaders that articulate their geopolitical interests in the AI race, be it in the form of "hard" or "soft" power. Despite their position in the world, the striving for military advantage and global regulation of AWS involves many other nations, especially Russia, Israel, South Korea, the UK, Australia, Germany and France. These countries also harbour companies that are leading in robotic military innovation and their governments actively engage in or are confronted with geopolitical tensions and conflicts.

As discussed above ("Approaching autonomous weapons embedded in sociotechnical imaginaries" section), sociotechnical imaginaries encompass broad concepts such as social order and nationhood. For this reason, the empirical material we refer to in the analysis necessarily reflects only a fraction of a multitude of cultural texts that fuel particular meanings of AWS. In this context, our objective is to specifically focus on those imaginations around AWS promoted in state military contexts and hence we pertain to two main discursive arenas: Firstly, the negotiation process at the CCW represents the international regulatory forum of

the UN, with talks taking place in Geneva since April 2013 [47]. Here, the USA and China have issued multiple position papers via the Group of Governmental Experts (GGE) on LAWS regarding the ongoing negotiations. They give their stance on definitional issues, the role of technical features and human intervention with a view to agreeing on a final and unanimously agreed upon UN protocol. The negotiations are still ongoing in 2022 and have been characterised by tedious definition struggles and gridlocks in the past. In a joint effort, Germany and France have proposed to conclude the CCW negotiations with a legally non-binding declaration [48], trying to mediate between two groups of countries that either strictly oppose a ban or call for effective and binding regulation [49]. With the recommendation of the 2019 GGE on LAWS, eleven guiding principles were adopted by the 2019 Meeting of the High Contracting Parties to the CCW. In 2021–2022, the CCW is aiming [12] to convert these voluntary principles into a "normative and operational framework" [50], but given that the CCW decision making requires consensus, it is estimated that "the probability of this forum producing a framework with unanimous agreement is very low" [51].

Secondly, we refer to position papers, directives, guidelines or decrees addressing AWS published by ministries, executives, higher secretaries or party assemblies of both nations that are publicly accessible[13]. National standpoints towards tech policy are not limited to one condensed official document or even one type of medium alone. Documents that receive the status of a strategy paper vary in medium and form of presentation, being themselves subject to differing political cultures. Clearly, China and the USA have different institutional traditions in announcing political agendas, due to opposing governmental systems and doctrines, e.g. CCP party rule in China vs. executive presidency in the USA. Further, these tech policy documents are not set in stone but are subject to substantive updates, adjustments or even radical dismissals and reorientations in light of new states of affairs in global politics, changes of ruling governments or the implementation of new doctrines. In sum, the empirical body (Table 1) comprises all relevant CCW standpoint papers of the USA and China that have been published since the start of the negotiations in 2013 and incorporates governmental

[12] During the completion of this paper in February 2022, these negotiations were still ongoing.

[13] Chinese papers are especially difficult to access. Also, the authors do not speak Chinese, so we limitted ourselves to official documents which depict an appropriate translation (thus, the papers that are deliberately directed to allies and adversaries, which suits the analytical agenda of this paper well).

documents addressing AWS [(or synonymously military (use of) AI)] since the year 2011, when the USA, as a first government, published a comprehensive DOD directive on autonomy in weapon systems (introduced in “Functional approaches to what “autonomous weapon systems” can and cannot do” section).

As a typology, the position papers offer various levels of analysis. First and foremost, the documents stemming from these two discursive arenas provide technical and definitional details on LAWS, showing many similarities to the academic debate (“The challenges of defining autonomous weapon systems” section). But beyond that, these position papers contain additional modi and layers of political communication. On the one hand, they act as self-assurances in the assessment of the current national security situation in the world and their own position in it. On the other hand, these documents can be instrumentalised to serve *realpolitik* interests. They set orientation points and geopolitical goals, identify threats and forge counter-strategies. Both countries are well aware of the signalling power of these documents for past, existing or emerging partners and adversaries. Further, apparently technical documents can offer strategic opportunities to escape definite LAWS regulation, or they can be used to deliberately provide a breeding ground for ongoing confusion in agreeing upon the regulatory object (see also “Technological definitions and normative understandings of AWS” section below).

Table 1 Overview of published CCW standpoint papers and governmental documents concerning LAWS of the USA and China 2011–2022

Country	Name of Document	Date of Publication	Type of Document		Abbreviation
			CCW	Gov.	
United States of America	U.S. Opening Statement at the CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems	April 15, 2015	x		US.CCW1
	Characteristics of Lethal Autonomous Weapons Systems	November 10, 2017	x		US.CCW2
	Humanitarian Benefits of Emerging Technologies in the Area of Lethal Autonomous Weapons	March 28, 2018	x		US.CCW3
	Characterization of the systems under consideration in order to promote a common understanding on concepts and characteristics relevant to the objectives and purposes of the CCW	April 10, 2018	x		US.CCW4
	U.S. Statement on the Outcome of the GGE	April 13, 2018	x		US.CCW5
	Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems	August 28, 2018	x		US.CCW6
	Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems Geneva	March 28, 2019	x		US.CCW7
	CCW Meeting of High Contracting Parties – U.S. Statement on emerging issues	November 14, 2019	x		US.CCW8
	Documents Reflecting U.S. Practice Related to Emerging Technologies in the Area of Lethal Autonomous Weapons Systems	June 11, 2021	x		US.CCW9
	DoD Directive on Autonomy in Weapon Systems	November 21, 2012		x	US.PosP1
	Deputy Secretary of Defense Speech: The Third U.S. Offset Strategy and its Implications for Partners and Allies	January 28, 2015		x	US.PosP2
	Summary of the 2018 White House Summit on AI for American Industry	May 10, 2018		x	US.PosP3
	Unmanned Systems Integrated Roadmap 2017-2042	August 28, 2018		x	US.PosP4
	Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity	February 12, 2019		x	US.PosP5
	Maintaining American Leadership in Artificial Intelligence	February 14, 2019		x	US.PosP6
	DoD Digital Modernization Strategy	July 12, 2019		x	US.PosP7
	The United States Air Force Artificial Intelligence Annex to The Department of Defense Artificial Intelligence Strategy	September 12, 2019		x	US.PosP8
	Ensuring American Leadership in Automated Vehicle Technologies. Automated Vehicles 4.0	December 23, 2019		x	US.PosP9
	National Security Commission on Artificial Intelligence	July 13, 2020		x	US.PosP10
	Executive Summary: DoD Data Strategy. Unleashing Data to Advance the National Defense Strategy	September 30, 2020		x	US.PosP11
	Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems	December 01, 2020		x	US.PosP12
	U.S. Department of Homeland Security Artificial Intelligence Strategy	December 03, 2020		x	US.PosP13
	Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems (updated)	November 17, 2021		x	US.PosP14
	International Discussions Concerning Lethal Autonomous Weapon Systems	December 21, 2021		x	US.PosP15
	Emerging Military Technologies: Background and Issues for Congress	April 06, 2022		x	US.PosP16

AWS as geopolitical signifiers: strategies in political communication in China and the USA

China and the USA employ different strategies to put their AI-driven military dominance on display. Matter-of-fact tech policies and national strategies alternate with messages of

national superiority. This section focuses on this particular realm of political communication and employs a comparative analysis of both countries, dissecting how LAWS as AI imaginaries are employed as geopolitical signifiers of national particularities. It analyses them in terms of the military doctrines and AI imaginaries they promote ("Military doctrines, autonomous weapons and AI imaginaries" section) and the definitions of autonomous weapons they establish ("Technological definitions and normative understandings of AWS" section) which both cater to certain goals in political communication.

Military doctrines, autonomous weapons and AI imaginaries

Foreign geopolitics is embedded in military doctrines, serving as a signalling landmark for military forces, the reallocation of strategic resources and technological developments. The empirical material at hand offers layers of analysis hinting at national SIs that put AWS in broader frameworks. These frameworks inform the populace, allies and adversaries about national aspirations, while presenting military self-assurance as a tool to look into a nationally desired future (see "Approaching autonomous weapons embedded in sociotechnical imaginaries" section). Here, AWS act as an empty and hence flexible signifier, a proxy for a society that exhibits different national idealisations of social life, statehood and geopolitical orders.

Military doctrine: The United States of America

In January 2015, the Pentagon published its *Third Offset Strategy* [US.PosP2]. Here, the current capabilities and operational readiness of the US armed forces are evaluated in order to defend the position of the USA as a hegemon in a multipolar world order. The claimed military "technological overmatch" [ibid.], on which the USA's clout and pioneering role since the Second World War is based, is perceived as eroding. The Pentagon warns in a worrisome tone: "our perceived inability to achieve a power projection over-match (...) clearly undermine [sic], we think, our ability to deter potential adversaries. And we simply cannot allow that to happen" [ibid.].

The more recently published "Department of Defense Artificial Intelligence Strategy" [US.PosP5] specifies this concern with AI as a reference point. Specific claims are already made in the subtitle of the paper: "Harnessing AI to Advance Our Security and Prosperity". AI

should act as “smart software” [US.PosP5, p 5] within autonomous physical systems and take over tasks that normally require human intelligence. Especially, the US research policy targets spending on autonomy in weapon systems. It is regarded as the most promising area for advancements in attack and defence capabilities, enabling new trajectories in operational areas and tactical options. This is specified with current advancements in ML: “ML is a rapidly growing field within AI that has massive potential to advance unmanned systems in a variety of areas, including C2 [command and control], navigation, perception (sensor intelligence and sensor fusion), obstacle detection and avoidance, swarm behavior and tactics, and human interaction”.

Given that such ML processes depend on large amounts of training data, the DoD announced its Data Strategy [US.PosP11], harnessed inside a claim of geopolitical superiority, stating “As DoD shifts to managing its data as a critical part of its overall mission, it gains distinct, strategic advantages over competitors and adversaries alike” (p 8). In the same vein and under the perceived threat to be outrivalled, “the DoD Digital Modernization Strategy” [US.PosP7] lets any potential adversaries know: “Innovation is a key element of future readiness. It is essential to preserving and expanding the US military competitive advantage in the face of near-peer competition and asymmetric threats” [US.PosP7, p 14]. Here, autonomous systems act as a promise of salvation of technological progress, which is supposed to secure the geopolitical needs of the USA.

Specified with LAWS, the US Congress made clear: “Contrary to a number of news reports, U.S. policy does not prohibit the development or employment of LAWS. Although the USA does not currently have LAWS in its inventory, some senior military and defense leaders have stated that the USA may be compelled to develop LAWS in the future if potential US adversaries choose to do so” [US.PosP12, p 1].[14]

Remarkably, the USA republished the very same Congress Paper in November 2021, just by a minor but decisive alteration. It changed “potential U.S. adversaries” into “U.S. competitors” [US.PosP14]. While it remains unmentioned (and presumably deliberately so) who is meant

[14] Given the definition of LAWS, the USA’s claim of not possessing any LAWS is highly debatable. Such will be further discussed in Military Doctrine: China, looking at technical LAWS definitions.

by both “senior military and defence leaders” and so named “U.S. competitors”, this minor change hints at a subtle but carefully orchestrated strategic tightening of rhetoric, sending out the message that the US acknowledges a worsening in the geopolitical situation with regard to the AWS development. In reaction, the USA continue to weaken their own standards for operator control over AWS in the most recent 2022 Congress Paper (as of May 2022), reframing human judgement: “Human judgement [sic!] over the use of force does not require manual human “control” of the weapon system, as is often reported, but instead requires broader human involvement in decisions about how, when, where and why the weapon will be employed” [US.PosP16]. Certainly, the rhetorical “broadening” of the US direction lowers the threshold to employ AWS in combat, evermore distancing the operator from the machine.

This stands in stark contrast to the US position in earlier rounds of the CCW process; here, the USA not only claims that advancements in military AI are of geopolitical necessity but also portrays LAWS as being desirable from a civilian standpoint, identifying humanitarian benefits: “The potential for these technologies to save lives in armed conflict warrants close consideration” [US.CCW3, p 1]. The USA is listing prospective benefits in reducing civilian casualties such as help in increased commanders’ awareness of civilians and civilian objects, striking military objectives more accurately and with less risk of collateral damage, or providing greater standoff distance from enemy formations [US.CCW3]. Bluntly, the USA tries to portray LAWS as being not only in accordance but being beneficial to International Humanitarian Law and its principles of proportionality, distinction or indiscriminate effect (see also “Technological definition: United States of America” section). While such assertions are highly debatable and have been rejected by many [1, 5, 7, 8], they do shed a very positive light on military technological progress, equating it with humanitarian progress.

In a congress paper on AWS, published in December 2021, these humanitarian benefits are once more mentioned but only very briefly, while a sharpening of the rhetoric is clearly noticeable. The paper also summarises the CCW positions of Russia and China, implicitly clarifying who is meant by “U.S. competitors” (see above). China, even though only indirectly, is accused by invoking that “some analysts have argued that China is maintaining “strategic ambiguity” about its position on LAWS” [US.PosP15, p 2]. This is the first time the USA overtly

expresses in a position paper that it understands the AWS negotiations as a political power play, instead of serving the aim of finding an unanimously agreed upon regulatory agreement.

In sum, the USA claims a prerogative as the dominant and legitimate geopolitical player in a multipolar world order, who is under external threat. The ability to defend military supremacy against lurking rivals is portrayed as being in a dependent relationship with the level of technological development of the armed forces, specified with LAWS. The USA claim to hegemonial leadership may only be secured through maintaining technological superiority.

Military doctrine: China

The doctrinal situation in China is more complex and ambivalent. In 2003, the Chinese Communist Party (CCP) and the People's Liberation Army (PLA) announced the concept of the "Three Warfares", a military guideline for enforcing Chinese geopolitical interests that has been systematically embedded in the PLA's military doctrine in recent years [52]. This concept promotes the objective of framing key strategic arenas of foreign policy in one's favour, so that kinetic (physical military) interventions appear irrational to opponents. This framing, also known as "information warfare" [53], insinuates that international conflicts are less decided by armies carrying off the victory but rather by the media narratives that have the upper hand in interpreting the events.

The concept of "Three Warfares" has been discussed by numerous authors [52,53,54,55,56], encompassing the following dimensions: the so-called *psychological warfare* aims to influence or disrupt an opponent's ability to make decisions. This includes practices that deter, shock or demoralise competitors. Media warfare, on the other hand, aims at influencing and manipulating national and international public opinion in order to generate support for China's military interventions. This entails constant and insistent media exposure, which aims to influence the perception and attitudes of the domestic or enemy population. The third dimension focuses on the legal dimension ("lawfare"). Creative distortions and omissions, conceptual vagueness and loopholes in regulations and international legal conventions serve the purpose of expanding one's own operational possibilities while simultaneously thwarting opponents in their scope of action. This instrumentalisation of the legal framework should be understood as a means of a "rule by law not rule of law" [54].

The strategic orientation of the “Three Warfares” also reflects a concession to the current military and geopolitical supremacy of the USA. While the USA claims its global leadership with rhetorical boldness, China sketches a military SI of an “underdog”, focussing on tactics of asymmetric warfare. This enables it to avoid direct military confrontation on all fronts and deploy a policy of “shashoujian” (杀手锏), which should be translated as “trump-card” approach [57,58,59]. Instead of competing in all strategic arenas with the USA, this doctrine targets a selective approach, fostering military technology that “the enemy is most fearful of”, including the call that “this is what we should be developing” [60].

However, in recent strategy papers, China has presented itself more confidently. As with the US, AI now plays a crucial role as a “cutting-edge” technology in China’s foreign policy aspirations [61,62,63,64,65].

The AlphaGo win over professional Go player Lee Sedol in 2016, which received a lot of media attention in China (280 million live viewers) was coined by some authors a Chinese “Sputnik moment” [66, 67], hence a wake-up call, which may well have contributed to the massive increase in spending in tech industry and research. Certainly, with the 2017 “new generation artificial intelligence development plan” the CCP also embraces these bold AI ambitions rhetorically by emphasising the need to “grasp firmly the strategic initiative of international competition during the new stage of artificial intelligence development [and] create new competitive advantage” [CH.PosP4, p 2]. The CCP decisively calls for a technological superiority that is equipped “to build China’s first-mover advantage in the development of AI” [CH.PosP4, p 1].

Such new confidence and ambitions are similarly met with a multilateralist appeasement and peacekeeping positioning [CH.PosP9]. China claims full sovereignty and strict non-interference in questions of national interest and security. This relates to, among other things, the one-China unification principle (e.g. directed to Taiwan “China must be and will be reunited”) or territorial claims (e.g. “safeguard China’s maritime rights and interests”). Beyond this sphere of the national interest the CCP pictures a military SI of a global hegemon without expansive aggressions (“Never Seeking Hegemony, Expansion or Spheres of Influence”). Sources of instability are located elsewhere, namely, in local “separatism” and foreign aspirations with “order [...] undermined by growing hegemonism, power politics,

unilateralism and constant regional conflicts and wars". At the same time, the USA is blamed directly for posing a threat to "global strategic stability" [CH.PosP9].

In sum, China's military SI depicts a global player that has caught up on its rivals at a military level. The CCP adjusts its doctrines and strategies pragmatically, from an underdog position to an assertive hegemon, clearly addressing geopolitical claims and means to get there. Military doctrines are clearly linked, as with the USA, to modernist narratives of technological progress, incorporating intelligent weaponry as AWS as a means to an end to outrival competitors. The technological race for supremacy in this key strategic technology is perceived as open, with China claiming legitimate ambitions.

Technological definitions and normative understandings of AWS

The USA and China have published national strategy papers as well as position papers at the CCW that are of a technical nature, aiming to define AWS. These documents have to be read against the backdrop of the larger SIs as introduced above ("Approaching autonomous weapons embedded in sociotechnical imaginaries" section), motivating and legitimating the state's strategic interpretative flexibility in creating and promoting AWS definitions. Hence, these documents not only inform which understanding—and technological variation—of autonomous weapon systems is to be prioritised, but further raise the question to what greater ends these specific interpretations are pursued. For example, in much the same way as the US American definitions of AWS, the Chinese "lawfare objectives" keep the backdoor open for developing automated weapons that escape the poor attributions of autonomy found in the AWS documents, with many military applications remaining legally and politically unaffected. A closer look at the national AWS definitions in the following sections will illuminate this issue.

Technological definition: United States of America

The DoD Directive 2012/2017 [US.PosP1, emphasis added] provides seemingly unequivocal definitions:

> Autonomous weapon system. Targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are

> designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation."
>
> (...)
>
> Semi-autonomous weapon system. A weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator."

A first problem with the US definition arises with the role of the *human operator* as a defining *criterion for autonomy*. As discussed in "Definitions focusing on the degree of human control over supposedly autonomous systems" section, conceptually, the USA advocates a relational approach to autonomy, linking it to the human presence. But the essential question of what an autonomous system comprises cannot simply be addressed by determining whether a human is in the loop or not. The degree of human intervention may give us advice on how to *use* such weaponry, but it does not help much in defining what it *is*. As Crootof clarifies: "If a weapon system has the capacity to independently select and engage targets, whether there is a human supervisor or whether it is operated in a semi-autonomous mode is a question of usage—and thus regulation—and not of autonomy" [11]. Very powerful weapons can be controlled by an operator and restrained such that their fire power (e.g. operational speed, fire range or power of devastation) is actually rarely fully in use. But from this observation, we can hardly deduce that we have arrived at the very essence of what the weaponry actually is and what it is capable of. While the role of human intervention in AWS is ethically and politically a much-needed debate, but not a debate without pitfalls as discussed by various authors regarding "meaningful human control" [24,68,69,70,71,72], it simultaneously raises further confusion if it is regarded as an appropriate characteristic in defining AWS.

More problematically, making a definition of AWS dependent on human intervention creates new loopholes in escaping effective legal regulation. The fundamental problem with the DoD definition stems from the fact that the standards for autonomy are simply very low—actually, it does not do justice to the term autonomy at all. The definition does not engage with the complexity of the term, clarifying what is really meant by autonomy. Should autonomy be rather understood as *self-sufficiency*, or as *self-directedness*, and hence as independence from outside control [73] (see "Technical definitions of autonomy and autonomous weapons systems" section)? Also, as problematised above, operation under pure autonomy as the DoD

document suggests is a myth, as any technical device is influenced by external factors such as technical infrastructure, terrain etc.

In essence, the DoD reduces the term autonomy to a process of *automation:* Any (non-) trivial system—either mechanical or algorithm-based—that, once activated, automatically processes (hence, without further human intervention) tasks and interacts with an environment would meet this criterion. Following the US reasoning, it is extremely hard to differentiate between advanced and very rudimentary mechanical or algorithmic systems, as literally *any* of them can be reduced to processes of automation. Thus, reducing autonomy to a process of automation introduces the notion of a continuum, making a clear differentiation between ubiquitously labelled "intelligent" weaponry impossible and the distinction between full or only semi-autonomy ever more complicated (cf. "Definitions focusing on the degree of human control over supposedly autonomous systems" section).

Take, for example, the case of radar detection systems, which have been in use for decades and which are capable of *identifying*, *selecting* and *targeting* enemy objects without the necessity for human intervention. The only difference between such systems and AWS would be the capability of automatically engaging with these targets. But weapon systems that fulfil such additional criteria have existed for years already, with the best example maybe being the Phalanx system [74][15], which has been in use since the 1980s, and hardly raised any regulatory concern back then [75]—especially not from the US side.

Problematically, the DoD definition cannot account for military advancements in fire power or complex machine behaviour such as *adoption* enabled through new data processing capabilities in machine learning—leading to a new myriad of problems such as *unpredictability* [76, 77] or *opacity* [78, 79] of machine behaviour, which are connected to safety, incomprehensibility and accountability issues well known from the civil AI regulatory debate. These phenomena in turn raise the fundamental question of whether deploying LAWS violates the Geneva Convention of IHL. If machine behaviour becomes ever more

[15] "Close-in Weapon Systems (…) designed to engage anti-ship cruise missiles and fixed-wing aircraft at short range. Like other close-in weapon systems, Phalanx provides ships with a terminal defense against anti-ship missiles that have penetrated other fleet defenses. (…) Unlike many other CIWS, which have separate, independent systems, Phalanx combines search, detection, threat evaluation, acquisition, track, firing, target destruction, kill assessment and cease fire into a single mounting" [74].

unpredictable, opaque and complex, it is debatable if the Geneva principles of the IHL *distinction*, *proportionality* and *accountability* in *hors de combat* can be met at all [80,81,82].

The USA has never claimed to retain from developing LAWS; in fact, it even cherished its advantages (see "The United States of America" section [US.CCW3]) and, as discussed above, threatens adversaries to "develop LAWS in the future if US competitors choose to do so" [US.PosP15]. This statement is, if one takes the DoD definition as a reference, strictly speaking, false. As discussed in relation to the Phalanx system, the USA have used LAWS in the past already and still do so today [16][Us.PosP12] [83, 84].

Conclusively, the DoD definition has the problematic effect of levelling down so many weapon systems under one category that critical advancements in weapon abilities that are now underway cannot be accounted for (making compliance with the Geneva principles more challenging). With such a vague and all-encompassing definition, effective legal regulation is ever more complicated, ensuring that national advances in the development of LAWS are not impeded.

Technological definition: China

China's contributions to the discussions at the CCW are rather limited, but serve well to understand China's ambivalent stance on AWS, echoing its international normative positioning (as introduced in "Military Doctrine: China" section). Their ambiguity helps to keep a strategic backdoor for optionality open. In the 2017 CCW negotiations, China adopted a positive stance on international regulation, favouring preventive arms control: "The international community should follow the concept of universal security on the basis of existing international law, carry out preventive diplomacy, check the trend of an arms race in the high-tech field and maintain international peace and stability" (12th December 2017, p 5). This is in accordance with the multilateralist stance voiced in the general AI policy trajectory of the country ("Actively participate in global governance of AI (...), Deepen

[16] For example, the so-called fire-and-forget weaponry such as the LRASM stealth anti-ship cruise missile in the US arsenal which can travel around 500 nautical miles before hitting target. But the DoD directive [US.PosP1] and the Congressional Research Service to the US congress label such weapon types solely "semi-autonomous", justified by humans doing the target selection through "autonomous functions" [Us.PosP12]. Such labelling clashes with many other experts in the field who categorise these weapons as autonomous [69, 75].

international cooperation in AI laws and regulations, international rules (...) and jointly cope with global challenges" [CH.PosP4, p 25] [85]).

Such a preventive regulatory stance was regarded more critically in 2018. Here, China states that "(...) the impact of emerging technologies deserve objective, impartial and full discussion. Until such discussions have been done, there should not be any pre-set premises or prejudged outcome, which may impede the development of AI technology" [CH.CCW2, p 2]. This rather innovation and military friendly policy reveals clear reservations against a precautionary principle that would regulate LAWS restrictively and prevent an AI arms race. The ambivalence seems even more striking when looking at the Chinese LAWS definition presented at the CCW:

Definition [CH.CCW2, p 1, enumeration added by authors for better overview]

According to the Chinese view, "LAWS should include but not be limited to the following 5 basic characteristics": (1) Lethality, "which means sufficient pay load (charge) and for means to be lethal"; (2) Autonomy, "which means absence of human intervention and control during the entire process of executing a task"; (3) Impossibility for termination, "meaning that once started there is no way to terminate the device"; (4) Indiscriminate effect, "meaning that the device will execute the task of killing and aiming regardless of conditions, scenarios and targets"; (5) Evolution, "meaning that through interaction with the environment the device can learn autonomously, expand its functions and capabilities in a way exceeding human expectations".

Conceptually, these LAWS criteria display a pick-and-mix approach, with the first stating the obvious, with the second showing strong similarity to the US definition (with its discussed pitfalls), with the fourth showing compliance to the Geneva Principles of IHL, and with the fifth hyperbolising, picking a fancy term "evolution" (hence lending imagination from a biological domain and maybe even evoking fantasies of an organic, autopoetic and reproductive machinery creating awe by exceeding human capabilities) to label *adoption* in machine learning processes.

The real crux lies in the third of these criteria, which hypothesises that once started, there is no way to terminate a device. In essence, this scenario describes a universally destructive,

actually ludicrous idea, which is nothing but absurd. Machines are not *perpetuum mobiles* but rely heavily on infrastructure, supervision, context, etc.—so, clearly, machinery self-sufficiency is a myth (see "Technical definitions of autonomy and autonomous weapons systems" section). Strictly speaking, these criteria depict sensational doomsday fiction, once more proving the hybridity of the entire AWS discourse, where *realpolitik*, imagination, possibility and fiction are conflated [86][17] ("Approaching autonomous weapons embedded in sociotechnical imaginaries" section).

It is exactly these unrealistic criteria for autonomous weapons that maintain the idea of promoting seemingly less dangerous—only "automatic"—weapon systems, undermining national or international legislation efforts. Where the US definition has set the benchmark for AWS too low, the Chinese set the benchmark for AWS too high, rendering their existence near science fiction. Hence, demands to ban AWS following these criteria can largely be understood as a political gesture of purely symbolic value. Implicitly, the development of autonomous and semi-autonomous weapon systems is not only tolerated but by definition appears as a legitimate course of action. This perfectly voices the objectives laid out in so-called asymmetric lawfare (see "Military doctrines, autonomous weapons and AI imaginaries" section): The legally vague, even bland criteria applied in the description and definition of LAWS have the intended effect of not curtailing one's own political scope of action.

Conclusively, both countries are against a complete ban on AWS, and with the definitions they promote at the CCW, they certainly do leave a backdoor open for further development and use.

Conclusion

This paper reveals the ways in which (lethal) autonomous weapon systems (AWS) are used as flexible reference objects in political communication. It shows how the USA and China embed AWS in their military doctrines and uncovers idealisations of geopolitical orders. The analysis navigates between different theoretical disciplines in order to deconstruct these national

[17] The German Delegation went even further into the science fiction genre, blunlty alleging: "Having the ability to learn and develop self-awareness constitutes an indispensable attribute to be used to define individual functions or weapon systems as autonomous" [86].

quests, which are interpreted as competing sociotechnical imaginaries (SIs). Both nations employ semantic manoeuvres in the realm of LAWS to enforce their military interests. The chosen approach—which involved considering AWS as geopolitical signifiers of national particularities—reveals both similarities and differences. This is hardly a surprise, since SIs are strategically deployed as part of political communication: only by making the motifs mutually decipherable while at the same time stressing differences can both sides ensure an intelligible back and forth in communication.

The main objective shared by both sides is the attempt to cater to certain goals in political communication. In particular, the two nations use the term AWS as a semantic means of deterrence in hybrid warfare. More recent political developments illustrate an escalating rhetoric that also points to the function of military technology as a semantic vessel. On the US side, subtle terminological changes (such as substituting “potential U.S. adversaries” for “U.S. competitors”) have been accompanied by an increasingly transparent and conscious unmasking of the CCW negotiations as an arena of rhetorical contest. The worsening of the international security situation has motivated the USA to lower its standards of human control over AWS, which makes the employment of AWS more likely. Such endeavours are undermining international humanitarian efforts at establishing binding and supranational rules to regulate AWS. On the Chinese side, the doctrine of overt lawfare and media warfare have been obvious since the PLAs announcement in 2003. Recently, this self-portrayal has painted the picture of a transformation from an “AI underdog” to an assertive hegemon by means of AI superiority.

In another conspicuous similarity, the military doctrines of both countries are clearly linked to narratives of technological progress, with the USA and China emphasising that intelligent weaponry can be used to safeguard their respective geopolitical goals (especially regarding disputed territories and spheres of influence). AI technologies are tied to overt efforts to enforce legitimacy for military technology advancements and aggressive military strivings. Technological superiority is elevated to a sublime status and portrayed as indispensable to secure national orders in a perceived arena of fierce international competition (AI weapons race). The emphasis of national resilience to defend military hegemony (US), or to catch up and achieve a pole position (China), brings to the fore larger national imaginaries that

articulate idealisations of world orders and their respective value foundations. AWS informed by SI, especially in a broader context of AI, articulate visions of national pride that are sought in technological advancement and achievement, even if at times they are hidden behind the smokescreen of international collaboration.

Major differences are apparent in the linguistic manoeuvres by which the USA and China achieve their goals. The US military definitions of AWS—which are also a conceptual blueprint for many other institutions and organisations—operate on a conceptual continuum, mainly reducing autonomous qualities to processes of automation. Taken together with the relational understanding of autonomous systems (which always necessarily involves human agency), this effectively creates a hybrid understanding of automatic and/or autonomous (weapons) systems. This blurring makes it all the more challenging to find legal parameters for the regulation of AWS. As an effect of this indeterminacy, national ambitions with regard to the development of novel weapon technologies remain unaffected: this lack of clarity allows for a historical perspective, focusing on functions such as target selection and engagement, which draws a continuous line from CIWS systems to today's elaborated systems. Innovative technological features, which include machine learning operations and for this reason enable unprecedented adaptive qualities and unpredictable behaviour, remain largely unaccounted for in the AWS definition by the USA.

The understanding of AWS promoted by China at the CCW has intentionally fostered an ambiguity in defining AWS that helps to keep the strategic backdoor for the development of "intelligent" weapons open, despite the publicly displayed efforts to curtail their development and use. This is on the one hand achieved by taking an ambivalent stance to preventive measures against novel technologies and on the other by promoting a wildly contradictory and bizarrely unrealistic understanding of AWS. It is the latter in particular that helps to legitimise the use of automatic weapons, which are indirectly portrayed as the much less worrisome technology.

On an international level, the semantic ambiguities of both states, which employ value-laden concepts such as machine *autonomy* and (human) *control* in the context of AWS, are deliberately exploited in order to usurp efforts for their effective regulation. Effectively, both nations are undermining global efforts to prevent an AI weapons race—even if they are

simultaneously promoting a rhetoric of appeasement and collaboration. If autonomous systems are understood as a relational quality that is always interwoven with external factors, the difference between them and "only automatic" systems is blurred. This means that novel military technologies seem fully legitimate as they are presented as a mere continuation of the weapon systems of the past, which did not spark a lot of controversy back then. If, on the other hand, autonomy and autonomous systems are defined as entities that operate completely independently of external factors such as infrastructure, energy supply, human oversight or decisions, the portrayal of AWS crosses the boundary into the realm of what is conceptually impossible. Regulating AWS becomes a vain endeavour since these technologies do not exist. In an effort to undermine much needed international regulation, it is exactly this paradoxical double-bind that ensures that states can continue the development of highly automatic and destructive weaponry.

The European actors have not contributed to an effective regulation of LAWS either. Neither Germany nor France as powerful EU nations are listed as countries that call for a prohibition on fully autonomous weapons by the Campaign to Stop Killer Robots, even though they are both active in the CCW process [87]. Their efforts for a voluntary regulatory framework can be perceived as less affirmative than other countries that strictly oppose a ban on LAWS, but this just seems to be another manoeuvre to circumvent tight regulation. The USA has happily exploited the German and French initiative as a model for "alternative approaches to manage LAWS" and is now advertising its own "nonbinding Code of Conduct to "help States promote responsible behaviour and compliance with international law" [US.PosP15]. Effectively, these declarations should be understood as a fig leaf strategy that mobilises a more humane rhetoric while striving for legitimacy for a soft LAWS regulation approach.

From a theoretical and analytical standpoint, a multidisciplinary lens is pivotal in the effort to make sense of the complex interdependence of conceptual frameworks, technological applications and a performative rhetoric. This lens also significantly sharpens our understanding of how they contribute to the present and future development of weapons technologies and the meanings attributed to them. It has the potential to inspire much needed research on the different political, legal and cultural (semio)spheres to further illuminate the functions and effects of AWS embedded in SIs.

When such momentous technologies are at issue, it is of paramount importance to defend the valence of concepts such as *autonomy*, *accountability* and *responsibility*. It is an imperative to prevent these values from being watered down as a consequence of power plays in the political arena.

References

1. Bhuta N, Beck S, Geiß R, Liu H-Y, Kreß C (eds) (2016) Autonomous weapons systems: law, ethics, policy. Cambridge University Press, Cambridge.
2. Krishnan A (2009) Killer robots: legality and ethicality of autonomous weapons. Ashgate Publishing, Burlington.
3. Scharre P (2018) Army of none: autonomous weapons and the future of war. W. W. Norton & Company, New York.
4. Ernst C (2019) Beyond meaningful human control? – interfaces und die imagination menschlicher Kontrolle in der zeitgenössischen Diskussion um autonome Waffensysteme (AWS). In: Thimm C, Bächle TC (eds) Die Maschine: Freund oder Feind? Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-22954-2_12 Chapter.
5. Article36. https://article36.org. Accessed 14 Sept 2021.
6. Campaign to Stop Killer Robots. https://www.stopkillerrobots.org. Accessed 14 Sept 2021.
7. Future of Life Institute (2015) Autonomous weapons. An Open Letter from AI & Robotics Researchers. https://futureoflife.org/open-letter-autonomous-weapons. Accessed 14 Sept 2021.
8. International Committee for Robot Arms Control (ICRAC). https://www.icrac.net. Accessed 14 Sept 2021.
9. Jasanoff S (2015) Future imperfect: science, technology, and the imaginations of modernity. In: Jasanoff S, Kim SH (eds) Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. University of Chicago Press, Chicago/London, pp 1–33.
10. Jasanoff S, Kim SH (eds) (2015) Dreamscapes of modernity: sociotechnical imaginaries and the fabrication of power. University of Chicago Press, Chicago/London.
11. Crootof R (2015) [2014] The killer robots are here: legal and policy implications. Cardozo L. Rev. 36(1837-1915):1854–1862.
12. Christman J (2018) Autonomy in moral and political philosophy. In: The Stanford Encyclopedia of Philosophy (Spring 2018 Edition). Center for the Study of Language and Information (CSLI). Stanford University. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/. Accessed 14 Sept 2021.
13. Khurana T (2013) Paradoxes of autonomy: on the dialectics of Freedom and normativity. Symposium 17(1):50–74. https://doi.org/10.5840/symposium20131714 Article.
14. Rebentisch J (2012) Aesthetics of installation art. Sternberg Press, London
15. Bradshaw J, Hoffman R, Woods D, Johnson M (2013) The seven deadly myths of "Autonomous Systems". IEEE Intelligent Syst 28:54–61 pp 2–3.
16. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam, p 59.
17. United Nations (2021) Background on LAWS in the CCW. https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/. Accessed 30 June 2021.

18. Lang J, van Munster R, Schott RM (2018) Failure to define killer robots means failure to regulate them. States disagree on definition of lethal autonomous weapons, DIIS Policy Brief. https://www.diis.dk/en/research/failure-to-define-killer-robots-means-failure-to-regulate-them. Accessed 14 Sept 2021.
19. Noorman M, Johnson DG (2014) Negotiating autonomy and responsibility in military robots. Ethics Inform Technol 16(1):51–62. https://doi.org/10.1007/s10676-013-9335-0.
20. Sauer F (2016) Stopping 'killer robots': why now is the time to ban autonomous weapons systems. Arms Control Today 46(8) https://www.armscontrol.org/act/2016-09/features/stopping-%E2%80%98killer-robots%E2%80%99-why-now-time-ban-autonomous-weapons-systems. Accessed 14 Sept 2021.
21. Schaub G, Kristoffersen JW (2017) In, on, or out of the loop? Denmark and Autonomous Weapon Systems. In: Centre for Military Studies' policy research. Centre for Military Studies. University of Copenhagen, Copenhagen https://cms.polsci.ku.dk/publikationer/in-on-or-out-of-the-loop/In_On_or_Out_of_the_Loop.pdf. Accessed 14 Sept 2021.
22. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam p 67.
23. International Committee of the Red Cross (2016) Autonomous Weapon Systems, Implications of increasing autonomy in the critical functions of weapons. Expert meeting, Versoix, Switzerland, p 8.
24. Böll Foundation (2018) Autonomy in Weapon Systems. The military application of artificial intelligence as a litmus test for Germany's new foreign and security policy, vol 49. Böll Foundation Publication Series on Democracy, Berlin, pp 20–21
25. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam p 70.
26. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam pp 74-76.
27. Reeves S, Johnson W (2014) Autonomous weapons: are you sure these are killer robots? Can we talk about it? Army Lawyer 1:25–31. https://ssrn.com/abstract=2427923.
28. Jasanoff S (2015) Future imperfect: science, technology, and the imaginations of modernity. In: Jasanoff S, Kim SH (eds) Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. University of Chicago Press, Chicago/London, p 4.
29. Sismondo S (2020) Sociotechnical imaginaries: an accidental themed issue. Soc Stud Sci 50(4):505–507. https://doi.org/10.1177/0306312720944753
30. Mager A, Katzenbach C (2021) Future imaginaries in the making and governing of digital technology: multiple, contested, commodified. New Media Soc 23(2):223–236. https://doi.org/10.1177/1461444820929321.
31. Kurzweil R (2005) The singularity is near. Viking Books, New York
32. Bostrom N (2014) Superintelligence. Paths, dangers, strategies. Oxford University Press, Oxford.

33. Bareis J, Katzenbach C (2021) Talking AI into being: the narratives and imaginaries of national AI strategies and their performative politics. Sci Technol Hum Values. 3. https://doi.org/10.1177/01622439211030007.
34. Natale S, Ballatore A (2017) Imagining the thinking machine: technological myths and the rise of artificial intelligence. Convergence 26(1):3–18. https://doi.org/10.1177/1354856517715164.
35. Beckert J (2016) Imagined futures: fictional expectations and capitalist dynamics. Harvard University Press, Cambridge, p 173.
36. Franklin HB (2008) War stars. The Superweapon and the American Imagination. University of Massachusetts Press, Amherst.
37. Singer PW (2010) Wired for War. The robotics revolution and conflict in the twenty-first century. Penguin Books, New York.
38. Lenoir T, Caldwell L (2018) The military-entertainment complex. Harvard University Press, Cambridge.
39. Maurer K, Graae AI (2021) Drone imaginaries: the power of remote vision. Manchester University Press, Manchester.
40. Baudrillard J (1995) The gulf war did not take place. Indiana University Press, Bloomington.
41. Singer PW, Brooking ET (2018) Likewar. The weaponization of social media. Eamon Dolan/Houghton Mifflin Harcourt, Boston.
42. Cummings ML (2018) Artificial intelligence and the future of warfare. In: Chatham House Report. Royal Institute of International Affairs, London, pp 7–18. https://euagenda.eu/upload/publications/untitled-209846-ea.pdf. Accessed 14 Sept 2021.
43. Newton MA (2015) Back to the future: reflections on the advent of a Autonomous weapons systems. Case Western Reserve J Int Law 47(1):5–23.
44. Coeckelbergh M (2011) From killer machines to doctrines and swarms, or why ethics of military robotics is not (necessarily) about robots. Philos Technol 24(3):269–278.
45. Bhuta N, Beck S, Geiß R (2016) Present futures: concluding reflections and open questions on autonomous weapons systems. In: Bhuta N, Beck S, Geiß R, Liu H-Y, Kreß C (eds) Autonomous Weapons Systems. Law, ethics, policy. Cambridge University Press, Cambridge, pp 347–374.
46. Geiß R (ed) (2017) Lethal autonomous weapons systems: technology, definition, ethics, law & security. Federal Foreign Office, Berlin.
47. Reaching critical will. https://reachingcriticalwill.org/disarmament-fora/ccw. Accessed 14 Sept 2021.
48. Group of Governmental Experts of the High Contracting Parties (2017) For consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS). Submitted by France and Germany, Geneva. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2017/gge/documents/WP4.pdf. Accessed 22 Feb 2022..
49. Delcker J (2018) France, Germany under fire for failing to back 'killer robots' ban. In: Anderlini J (ed) Politico. Axel Springer, Brussels (in press).
50. Group of Governmental Experts of the High Contracting Parties (2019) Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to

Be Excessively Injurious or to Have Indiscriminate Effects, Geneva. https://documents.unoda.org/wp-content/uploads/2020/09/1919338E.pdf. Accessed 14 Sept 2021..
51. Lethal AWS. Global Debate: what are countries doing about this issue? https://autonomousweapons.org/global-debate/. Accessed 14 Sept 2021.
52. Kania EB (2016) The PLA's latest strategic thinking on the three warfares. China Brief 16(13):10–14. https://jamestown.org/program/the-plas-latest-strategic-thinking-on-the-three-warfares/. Accessed 15 May 2020.
53. Timothy AW (2012) Brief on China's three warfares. In: Delex Special Report-3. Delex Consulting, Studies and Analysis (CSA), Delex Systems, p 4. http://www.delex.com/data/files/Three%20Warfares.pdf. Accessed 14 Sept 2021.
54. Halper S (2013) China: the three warfares. Prepared for Andrew Marshall, Director of the Office of Net Assessment, Office of the Secretary of Defence. https://cryptome.org/2014/06/prc-three-wars.pdf. Accessed 14 Sept 2021.
55. Jackson L (2015) Revisions of Reality. The three warfares—China's new way of war. In: Beyond Propaganda. Information at War: From China's Three Warfares to NATO's Narratives. The Legatum Institute, London, pp 5–15. https://li.com/wp-content/uploads/2015/09/information-at-war-from-china-s-three-warfares-to-nato-s-narratives-pdf.pdf. Accessed 14 Sept 2021.
56. Lee S (2014) China's 'three warfares': origins, applications, and organizations. J Strat Stud 37(2):198–221. https://doi.org/10.1080/01402390.2013.870071.
57. Allen G (2019) Understanding China's AI Strategy. Clues to Chinese strategic thinking on artificial intelligence and national security. In: Center for a New American Security https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy. Accessed 13 Mar 2019.
58. Bruzdzinski JE (2004) Demystifying Shashoujian: "China's Assassin's Mace" Concept. In: Scobell A, Wortzel L (eds) Civil-military change in china elites, institutes, and ideas after the 16th party congress. Diane Publishing Co, Darby, pp 309–364.
59. Kania EB (2020) "AI weapons" in China's military innovation. In: Global China. The Brookings Institution https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf. Accessed 15 May 2020.
60. Cheung TM, Mahnken T, Seligsohn D, Pollpeter K, Anderson E, Yang F (2016) Planning for innovation: understanding China's plans for technological, energy, industrial, and defense development, Report prepared for the US-China Economic and Security Review Commission, Washington DC, 28 July 2016. Citation of CMC Chairman Jiang Zemin, p 26.
61. Future of Life Institute (2018) AI policy - China. https://futureoflife.org/ai-policy-china/. Accessed 14 Sept 2021.
62. Horowitz MC (2018) Artificial intelligence, international competition, and the balance of power. Texas Natl Secur Rev 1(3):37–57. https://doi.org/10.15781/T2639KP49
63. Horowitz MC, Allen GC, Kania EB, Scharre P (2018) Strategic competition in an era of artificial intelligence. In: Center for a New American Security's series on Artificial Intelligence and International Security. Center for a New American Security. https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence. Accessed 14 Sept 2021.

64. Katzenbach C, Bareis J (2018) Global AI race: states aiming for the top. https://www.hiig.de/en/global-ai-race-nations-aiming-for-the-top/. Accessed 15 June 2019.
65. Roberts H, Cowls J, Morley J, Taddeo M, Wang V, Floridi L (2021) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI Society 36:59–77. https://doi.org/10.1007/s00146-020-00992-2
66. Kania EB (2017) AlphaGo and beyond: the Chinese military looks to future "intelligentized". warfare. https://www.lawfareblog.com/alphago-and-beyond-chinese-military-looks-future-intelligentized-warfare. Accessed 22 Feb 2022.
67. Lee K-F (2018) AI superpowers: China, silicon valley, and the new world order. Houghton Mifflin Harcourt, Boston, New York.
68. Crootof R (2016) A meaningful floor for "Meaningful Human Control". Temp Int'l Comp LJ 30:53–62.
69. Altmann J (2019) Autonomous weapon systems – dangers and need for an international prohibition. In: Benzmüller C, Stuckenschmidt H (eds) KI 2019: Advances in Artificial Intelligence. Joint German/Austrian Conference on Artificial Intelligence, Kassel, September 2019, Lecture Notes in Computer Science, vol, 11793. Springer, Cham, pp 1–17. https://doi.org/10.1007/978-3-030-30179-8_1
70. Amoroso D, Tamburrini G (2020) Autonomous weapons systems and meaningful human control: ethical and legal issues. Curr Robot Rep 1:187–194. https://doi.org/10.1007/s43154-020-00024-3
71. Chengeta T (2017) Defining the emerging notion of meaningful human control in weapon systems. J Int Law Politics 49(3):833–890.
72. International Committee for Robot Arms Control (2019) What makes human control over weapons systems 'meaningful'? Working paper submitted to the Group of Governmental Experts on lethal autonomous weapons of the. CCW, Geneva
73. Bradshaw J, Hoffman R, Woods D, Johnson M (2013) The seven deadly myths of "Autonomous Systems". IEEE Intelligent Syst 28:54–61 p 5.
74. NavWeaps. 20 mm Phalanx Close-in Weapon System (CIWS). https://doi.org/10.1177/1354856517715164. Accessed 14 Sept 2021.
75. Sauer F (2020) Stepping back from the brink: why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible. Int Rev Red Cross 102(913):235–259. https://doi.org/10.1017/S1816383120000466.
76. European Commission (2020) Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and Robotics. European Commission, Brussels.
77. Kowert W (2017) The foreseeability of human–artificial intelligence interactions. Texas Law Review 96(1):181–204.
78. Brkan M, Bonnet G (2020) Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: of black boxes, white boxes and fata morganas. Eur J Risk Regul 11(1):18–50. https://doi.org/10.1017/err.2020.10
79. Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 3(1). https://doi.org/10.1177/2053951715622512
80. Boulanin V, Bruun L, Goussac N (2021) Autonomous weapon systems and international humanitarian law. In: Identifying limits and the required type and degree of human–machine interaction. SIPRI Publications.

https://sipri.org/sites/default/files/2021-06/2106_aws_and_ihl.pdf. Accessed 13 Sept 2021.
81. Sassòli M (2014) Autonomous weapons and international humanitarian law: advantages, open technical questions and legal issues to be clarified. Int Law Studies 90(1):308–340.
82. Schmitt MN (2013) Autonomous weapon systems and international humanitarian law: a reply to the critics. Harvard Natl Sec J 4:1–37.
83. Department of the Navy (2019) Department of Defence Fiscal Year (FY) 2020 budget estimates. In: Justification Book Volume 1 of 1, Weapons Procurement. Navy. https://www.secnav.navy.mil/fmc/fmb/Documents/20pres/WPN_Book.pdf. Accessed 14 Sept 2021.
84. Vavasseur X (2021) NavalNews. Lockheed martin progressing towards LRASM integration on F-35. https://www.navalnews.com/naval-news/2021/01/lockheed-martin-progressing-towards-lrasm-integration-on-f-35/. Accessed 14 Sept 2021.
85. Kania EB (2018) China's strategic ambiguity and shifting approach to lethal autonomous weapons systems. https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethalautonomous-weapons-systems. Accessed 17 Sept 2021.
86. Permanent Representation of the Federal Republic of Germany to the Conference on Disarmament in Geneva (2018) Statement delivered by Germany on Working Definition of LAWS/"Definition of Systems under Consideration", Convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or to have indiscriminate effects, Geneva, p 2. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Germany.pdf. Accessed 14 Sept 2021.
87. Campaign to Stop Killer Robots (2020) Country views on killer robots. https://www.stopkillerrobots.org/wp-content/uploads/2020/05/KRC_CountryViews_7July2020.pdf. Accessed 22 Feb 2022.
88. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam, p 60.
89. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam, p 16.
90. Ekelhof MAC (2019) The distributed conduct of war: reframing debates on autonomous weapons, human control and legal compliance in targeting. Dissertation, Vrije Universiteit Amsterdam, p 17 Fn 15.
91. Hilgartner S, Miller CA, Hagendijk R (eds) (2015) Science and democracy. Making knowledge and making power in the biosciences and beyond, Routledge, New York/Abingdon.

ARTICLE IV

The realities of autonomous weapons: Hedging a hybrid space of fact and fiction[1]

Jascha Bareis & Thomas Christian Bächle

Abstract

The development of "Autonomous Weapon Systems" (AWS) has been subject to controversial discussions for years. They open up the hypothetical possibility of killing human populations without a human agent in the loop, which is the most extreme scenario that often lies at the centre of these debates. In order to draw a more complex picture, this publication project engages with the current social, political, cultural, ethical and military arenas of autonomous weapons in popular culture, regulatory debates, journalism and research. It adheres to an analysis of the different meanings articulated across these domains that constitute the "realities of autonomous weapons" and powerfully influence how we perceive and engage with these novel technologies. The articles in this volume analyse how the current debates on AWS mediate between fact and fiction and create a constant and complex dynamic between the actual technological developments and the potential futures that are associated with them. Paradoxically, it is exactly in this context of uncertainty – in which reality, imagination, possibility and fiction are conflated – that the full scope of this controversial technology becomes visible. Hence, the volume focuses on various practices, discourses and techniques in which AWS are both represented and created to become technological, military and political realities.

Keywords:

autonomous weapon systems, fiction, artificial intelligence, hybrid, mediations

[1] Introduction to the edited volume "The Realities of Autonomous Weapon Systems", publishing date spring 2025 with Bristol University Press und CC-BY license.

Introduction

The development of autonomous weapon systems (AWS) – at times also bearing the "lethal" label under the acronym LAWS – has been subject to intense discussions for years. Numerous political, academic or legal institutions and actors are debating the consequences and risks that arise with these technologies, in particular their ethical, social and political implications with many calling for a strict regulation, even a global ban. Despite this public prominence and the sheer consequentiality of these weapons, it often remains surprisingly unclear which technologies are evoked by the term AWS and what they are capable of. AWS can refer to drones, flight carriers, unmanned aerial, ground or maritime vehicles, robots and robot soldiers or cyber weapons such as computer viruses.

This uncertainty comes despite (or maybe because) of the fact that there are numerous definitions that try to specify the term either functionally ("once activated" autonomous weapons "can select and engage targets without further intervention by an operator", US Department of Defense, 2023: 21) or conceptually (derived from the theorisation of autonomous systems, artificial intelligence or machine learning). Definitions leave plenty of room for different types of technologies and – combined with the much wider discussions on AI – potentials and projections on future developments. Besides this terminological ambiguity, it also remains inherently vague in what sense and to what degree these systems can be characterised as autonomous at all. Even though the development of automated capabilities is undoubtedly advancing (Scharre, 2018), with an ever decreasing degree of human agency and ways to intervene, fully autonomous weapons that are completely beyond human control and for this reason feared by many, largely represent a conceptual possibility rather than an actual military reality.

From these ambiguities ensue wide gaps of meaning, which are in turn filled with imaginations – a common practice for new technologies and AI in particular (Suchman, 2023). Potential realities can fulfil an important role, as they are tools to transfer expert knowledge into other fields of society, including journalism, policy-making, research, education and democratic decision-making processes. Hence, the ideas on the functionality of AWS and their consequences are in this sense inspired and shaped by imaginaries on military, national and technological futures. They include geopolitical scenarios, ethical questions, national policies

or science-fiction. In security and military policies, these interconnections between different realities are even utilised as a methodology, for example as 'red teaming', which means applying creative fictional accounts of potential futures to inform actual decision-making (The Red Team, 2021). Another application is 'war gaming', a method of foreseeing future military scenarios originating at least in the 19th century but adapted to contemporary technological and media environments, including virtual reality and AI based simulations using large language models (Goecks & Waytowich, 2024).

The premise of AWS, seen as entertaining a hybrid space of their own, invites the exploration of the myriad of "realities of autonomous weapons". The rationale of the book maintains that the realities in question can only be understood by acknowledging the constant and complex dynamic between the actual technological developments and the visions and virtual scenarios that are associated with them. It is exactly in this context of uncertainty – in which imagination, possibility and fiction are conflated – that AWS become highly consequential. They provoke emotions, discourses, agitations, (re-)actions, investments, competition, policies or technological and military blueprints.

Publications on the topic of autonomous weapons often focus on their legal, political or ethical ramifications (e.g Bhuta et al., 2016; Krishnan, 2016), a first-order level of assessing these technologies, with some works also discussing their unique representations (Graae & Maurer, 2021). The foundation of these works is also based on the different realities sketched above. Introducing another way of analysing the "realities of autonomous weapons", the book puts forward a second-order level approach: an ethical problem, for example, is not framed only as such, that is along the lines of posing the question "which moral questions arise with automated killing machines?". The ethical problem, in the approach suggested here, is rather to understand it as a contributing factor that helps to construct, disseminate and maintain a specific understanding of lethal autonomous weapons in popular culture, politics, journalism or research. In short, ethical discourses co-create the realities of their object. For this reason, the perspectives taken in this book foreground the different realities of autonomous weapons and in turn aim at informing the existing debates about their (often implicit) underlying assumptions.

This introductory chapter of the book first outlines in more detail the conceptual approach taken here, discussing the theoretical underpinnings of the "realities of autonomous weapons". AWS are approached and understood as mediations, frictions and hybrid entities which create a reality of their own. They are theorised as both constitutive as well as performative to encompass the dynamics and different understanding they invoke around the globe. Subsequently, the chapter offers five reflections on these realities that help hedge and consolidate the dynamic meanings of autonomous weapons, which tend to receive so much attention in public, military and regulatory arenas. The chapter concludes with an overview of the book's structure and a brief summary of the individual contributions.

Approaching the realities of autonomous weapons

The idea of automatic or self-directed weapon systems can be traced far back (Galison, 1994). In military history, however, the final phase of the Cold War in the late 1980s and the Second Gulf War in 1991 can be seen as the key moment towards today's discourses on autonomous weapons systems, since it also saw the first philosophical examination of "intelligent" war machines (De Landa, 1991). Against the background of various ideas on "post-industrial" warfare (e.g. Echevarria & Shaw, 1992; Toffler & Toffler, 1993), the digitalisation of information and communication infrastructures of the US Armed Forces has been characterised as a "Revolution in Military Affairs" (Cohen 1996) and considered as a phase of disruptive technological developments. Around the same time the paradigm of "network-centric warfare" emerged, which defined the standards for a new form of warfare based on the idea to achieve permanent information dominance through rigorous networking of all forms of military systems, including both human and technical agents (see Ernst in this volume; Cebrowski, 2005).

Another milestone in the political and military ambitions to intensify the development of autonomous weapons systems – especially in the field of robotics – is marked by the terrorist attacks conducted on September 11, 2001 in the US and their aftermath (Singer, 2010). Most notably, weaponised drones such as the US MQ-9 "Reaper" (General Atomics) or the X-47 series (Northrop Grumman) were rapidly developed during a time that was labelled "War on Terror". Subsequently the notion of an "Age of Autonomous Systems" in warfare (Worcester,

2015) or calls to urgently start “preparing for war in the robotic age” (Work & Brimley, 2014) have emerged in recent years. Those visions were strongly driven by the military utilisation of more recent forms of AI such as machine learning tools or artificial neural networks (Cummings, 2017). The latest iteration of an innovative AI-related hype – at the time of writing – has been featured via the concept of “generative AI”, which has also entered both the vocabularies as well as the imaginations of military industries (Goecks & Waytowich, 2024).

The realities of autonomous weapons also include the dynamic between fact and fiction. They are often influenced by popular culture and inspired by more general assumptions about artificial intelligence and its relationships to the human in the broadest sense, echoing tropes such as the substitution of humans by machines, the risks of intelligent machines that are no longer subjected to human control. These realities are hence shaped by a mix of intentional framing and larger socio-cultural narratives that act on a discursive rather than an individual level and transcend the attribution of intentionality. A well-known position is the idea that autonomous weapons can be seen as more fair or just (Arkin, 2009). The obvious ethical and critical question is “What enables the framing of an instrument for surveillance and killing as an inherently ethical instrument? What kind of sociopolitical rationale underpins such a framing?” (cf. Schwarz, 2018). In other words, the framing of “ethicality” is produced by but also produces a particular realities of autonomous weapons.

The book also touches upon conceptual approaches to autonomous warfare technologies, shaping the ways in which they are modelled, developed or advertised in their interactions with humans. Well-known examples for this in the context of regulating autonomous weapons are the often normatively utilised descriptors of “meaningful human control” on the part of humans and “autonomous” on the part of machines. It is necessary to stress that both bear meanings that are constructed and constructive rather than descriptive (Bächle, 2023). These dynamic meanings prove to be particularly challenging in legal assessments that require a normative stance. Scholars have started to challenge the apparent consensus that "human judgement" is to be treated as legal requirements in the context of autonomous weapons. Querying the common foundations of arguments for AWS regulation – along the lines of explainability, accountability, dignity or the principle of humanity – and comparing AI-enabled technologies to other types of weapons, one question is still not settled: "If we want

better human control, we need to explain why" (Lecture held by Noam Lubell at the DILEMA conference in The Hague on 12 Octobre, 2023). Interestingly, this condition is not verbalised as strongly for other types of weapons systems (such as anti-personnel landmines), which can be equally harmful but are not met with a similar concern, involving explicit human oversight. The existing regulatory frameworks are seemingly sufficient in the case of less technologically advanced weapons. This is not to say that weapons of mass destruction (biological, chemical, radiological or nuclear) are any less controversial. Their development and actual employment, however, in most cases predates international regulatory frameworks (most notably that under the United Nations) and presents a different historical context. These weapons technologies, in other words, refer to both a wide array of legal and political histories and cultural representations. A technology like AWS, seen by many as genuinely novel, arguably triggers a heightened sense of uncertainty. Paired with the complexities of a multi-centered geopolitical context and competing media realities, the differing perception of urgency and threat – this is one of the book's assumptions – might in part be attributed to the fluctuating nature of the realities of autonomous weapons.

The "realities of autonomous weapons" are connected to – but not identical with – what Jasanoff and Kim (2015) call sociotechnical imaginaries. According to their definition, sociotechnical imaginaries are "collectively held, institutionally stabilised, and publicly performed visions of desirable futures, (...) and supportive of advances in science and technology" (Jasanoff & Kim, 2009: 120). Sociotechnical imaginaries inform realities of autonomous weapons especially in the field of state discourse and political communication, as communication in the public arena presupposes a shared understanding among larger social groups. In these public arenas imaginaries point to, as Jasanoff (2015) argues, "positive visions of social progress (...) [and], tacitly or explicitly, with the obverse — shared fears of harms that might be incurred through invention and innovation, or of course the failure to innovate" (Jasanoff, 2015: 4-5). The nationally shared visions and fears of our time are negotiated in light of major geopolitical shifts in the wake of the Russian aggression against Ukraine, looming conflicts between China, Taiwan and adjacent nations or the complex political situation in the Middle East. Depending on the respective point of view, AWS can be portrayed as a threat (losing against technologically superior adversaries) or as an opportunity or solution (to counter the problem of scarce human resources through automation).

However, the understanding of "realities" in this book goes further. The very idea of AWS is closely interwoven with military histories and current hopes and developments towards machine intelligence and the possibilities of human agency. Historically, AWS' military imaginations, contexts and discourses are continuous and dynamic developments that cannot be tied to one singular event or technical breakthrough. Rather, they can only be understood through the lens of their technical precursors and the shared norms and values of their time. The understandings that are associated with AWS also vary geographically, which means they cannot be reduced to one emblematic representation – often US and Euro-centric – such as killer robots or drone swarms (Coeckelbergh, 2011; Arquilla & Ronfeldt, 2000). The realities of autonomous weapons take into account popular aesthetics, fictions, policies and corporate discourses that can differ significantly cross-culturally.

This overlap between the technological paradigms, and their larger societal and cultural manifestations show that AWS are not only shared and understood in clearly articulated visions or imaginaries. They are characterised by mediation, frictions and hybridity that create a reality of their own. For example, efforts to predict future military threats, conflict scenarios and simulations under the condition of *potential* technological advancements is equivalent to the creation of 'as if' realities. These *virtual* – potentially innate – realities of autonomous weapons shape the *actual* debates on their ethical and legal ramification, the ways they are represented in public discourse and the basis of political decision-making today. For this reason, AWS are *created* as objects while at the same time drawing "distinctions between life and death, human and machine, culture and technology" (Karppi, Böhlen and Granata, 2018).

Media technologies have an important role in this (e.g. Hoskins & O'Loughlin, 2015), which is not limited to merely representing warfare and warfare technologies. Baudrillard famously commented that the Gulf War in 1991 was not taking place (Baudrillard, 1995). He described its reality as not bound to the battleground and constituted by actual combat operations – but as coming into effect via mediated, mainly televised form, broadcasting live into the living rooms of North American and European citizens. Mediatised and mediated warfare creates simulations of war, representations that do not presuppose actual events. The Gulf War points to the virtuality of war, it was not necessary for it to take place to become a reality in

the TV living rooms: a simulacrum in the Baudrillardian sense. The idea of mediated warfare became even more prevalent post 9/11: The paradigmatic importance of drones – in particular the claim of high-precision drone strikes – for the supposedly new forms of warfare is interrelated with normative questions associated with these weapons systems (Krasmann/Weber, 2015). From a technical standpoint, drones are not necessarily autonomous systems but rather remote-controlled robots (unmanned combat aerial vehicles), which are able to independently perform specific sub-tasks such as flying and reconnaissance. Nevertheless, drones have made a reality *imaginable*, in which technical autonomous systems are able to perform kill decisions independently from human control (Maurer and Graae, 2021). Their prominent representation in the media also established a particular aesthetics of drone images (see the work by Weilandt in this volume). A detached and distant view, reinforcing the narrative of technologically assisted clean forms of warfare against the enemy – favourably depicted as "terrorist vermin" in the 2000s (Sarasin, 2006). In a more abstract sense, drones have thus been established as both real technologies and symbols for the imagination of an expectable future, in which fully autonomous combat robots are no longer a purely fictitious possibility (Elish, 2018).

The mediated realities of AWS have to be accounted for, especially given new media environments, which incorporate virtual reality, augmentation and digital forms of decentralised communication – and lately, the rise of synthetically produced media with text and pictures through generative AI. This not only leads to a de facto convergence of military and entertainment media (Lenoir & Caldwell, 2018), when, for example, interfaces used to control drones are inspired and optimised by computer games and vice versa. But media forms themselves *shape* the realities of warfare, often in a fuzzy overlap of temporalities and media spheres. The recent conflicts in Ukraine and Gaza highlighted the ways in which social media publics are targeted in propaganda wars (Rudloff & Appel, 2023), while public authorities try to engineer opinions in a desired fashion. The new media environments also enable first person accounts of their experiences – evoking labels such as soldiers, terrorists, civilians, innocents – even allowing them to live-stream *their own* reality of on-the-ground combat (Rarm, 2023). It is impossible to ascertain whether these accounts are authentic or fake (Antinori, 2019).

Despite these vast fields of AI applications in hybrid warfare, and somewhat paradoxically, the public perception of autonomous weapons – promoted by state actors, the militaries or the industries – is often reduced to machinistic understanding of weapons: unmanned vehicles, missiles or drones. These materistical imaginations reduce the broad range of conducting attacks to an underestimated field of digital and AI-enabled warfare (Merrin, 2018; Shaw, 2016). Cyber attacks, however, quite holistically aim at the manipulation or destruction of computer software or devices, which disrupt not just militaries but potentially all aspects of our digital lives. 'Autonomous' computer viruses or cyber attacks do not just hit our capabilities to communicate, but potentially all mediated aspects of social reality and also the everyday material objects – "the Internet of Things" – that surround us (Arquilla, 2021). The manipulation of publics through misinformation, targeted leaks or the disruption of traditional media and journalism of media also thrives (Seib, 2021). The new media environment entails a power shift to platforms and private companies.

Acknowledging the tension, overlap and conflation of fact and fiction, the real and the virtual, the truthful and the fake, the desired and the detested, is the main conceptual baseline for the AWS case studies and analyses in this book. It is established (and good) practice for current research to strongly focus on normative issues of legal and ethical regulation of AWS in order to inform policy makers, politicians, the military industry and civil society. However, the "realities of AWS" takes a different, constructivist route to this end. It Interrogates different media, histories and visions, as well as geographical particularities for their realities. Thus, this volume aims to make explicit the tacit knowledge around AWS. It deconstructs their taken for granted preconditions and manifestations across the discourses and depictions of AWS and thereby, hopefully, is able to further substantiate the relevant normative debates with their legal and political implications.

The following five reflections are meant to pinpoint these complex realities of autonomous weapons by addressing common (mis)conceptions and by locating them within some of the larger contexts sketched above.

1. As autonomous weapons systems are perceived as clandestine technologies, their capabilities trigger curiosity and are often overestimated

AWS development is mostly classified. States conceal latest technology advancements in the name of national interest, with agencies and laboratories working on military innovations shielded from the public eye. Supremacy in weaponry power is trending high on many national and geostrategic security agendas (see for example Bächle & Bareis 2022 for a comparison of the US and China). It embodies a military and industrial striving for competitive advantage in a perceived arena of threat and rivalry. The urgency and legitimacy is derived from mobilising a rhetoric of fierce international competition, thereby hailing technological innovation as a pillar of national resilience capabilities (Bareis & Katzenbach, 2022).

A prominent example is the US Defense Advanced Research Projects Agency (DARPA). It was founded by president Eisenhower in 1958 and during its planning phase it was initially coined the "*Special Projects Agency*" (Barber Associates, 1975: 59). It was created in reaction to the Soviet induced Sputnik-shock. Still today its aim is to formulate and coordinate "breakthrough technologies and capabilities for national security" (DARPA, n.d.) together with academic research and industry. A self-assuring DARPA promotional video introduces the founding motif in 1958, which hails DARPA as being "the initiator, not the victim of strategic technological surprises" (DARPAtv, 2018: 0:24), catering to a rhetoric of fierce international competition and outrivaling.

Institutions like DARPA function as mission-oriented agencies (Mazzucato, 2011), which are legitimated by the imperative of state leadership often at the cost of democratic processes. It is common that they trade transparency and public accountability with speed and secrecy in the name of national interest. The role of public funding and the "hidden Developmental State" (Block, 2008) with agencies such as DARPA (or its European equivalent, the "Joint European Disruptive Initiative" (JEDI Foundation, n.d.) have changed throughout the years to become more similar to a network of public-private partnerships. State agencies cooperate with major technology corporations contributing to military and intelligence imperatives. Some of these projects were famously leaked in the past, such as the common surveillance practices by the US, made public by former intelligence employee Edward Snowden (Lyon, 2015). Or in the wake of the protest by Google against the plans to collaborate with the Pentagon under the name of project Maven, which in 2018 incorporated the company's AI technology in order to analyse drone surveillance footage (Simonite, 2021; see also Heffernan

in this volume). The idea of a hidden power structure gets also easily misused, for example by the Trump administration and in its aftermath by utilising a “deep state” conspiracy theory (Horwitz, 2022).

The concealing of state agencies and powerful companies in the name of “national interest” leaves imaginary space and rumour for the public and exploits a deep fascination with the inaccessible, clandestine - but seemingly powerful and out-of-control. Military industry and militaries’ showcase of weapons technology thrive in this context of uncertainty, as they can exploit public fascination and imagination, hailing the appearance or leaking of the suddenly novel and unprecedented. This fascination can be compared to the media rumble when the highly classified “Manhattan project”, the US research and development program from 1942 to 1946 of the nuclear bomb, lifted its curtain of secrecy. Overnight unknown scientists became showcased as national heroes in times of war and conflict. Technology became hailed as a means to rule the world and even heralded a new epoch of the anthropocene: the “nuclear age” (Hughes, 2004).

2. Autonomous weapons trigger both fascination and horror – and subscribe to common historical narratives of technology & dominance

The development and portrayal of AWS strongly speaks to and exacerbates the existing hopes and fears around AI (Cave & Dihal, 2019). Building on the age-old fascination for the latest technological development, they are simultaneously emblematic of potentially devastating effects and out-of-control scenarios playing with themes of dominance and control (see also Bode & Mohan investigating sentiments in the Indian public, or Jones analysing the stereotypes of female performing AWS in cinema history in this volume).

There are two historical narratives, one rotating around the concept of dominance, the other around enhancement and extension, that entertain sentiments of fascination and horror with technology. The first regards science and technology as ways to control and cultivate nature, essentially establishing both as distinct realms (Latour, 1993). Taming the natural environment and its unpredictable force (through droughts, floods or earthquakes) rationalises technology as a necessary force to expand and maintain human civilisation

through domination (Nye, 2004). Industrialisation and engineering projects such as the construction of dams or railway networks epitomise the "technological conquest of matter" (Marx, 2000, 197). Overcoming the physical limits of nature and matter plays on the imagination of achieving the seemingly impossible (Beckert, 2016).

The second historical discourse more directly refers to contexts of military technology as forms of enhancement and extension in an array of different techniques. Foremost, this refers to weapons technology which allows to increase the distance between soldiers, and also decrease the need to engage in direct body combat, including swords, bow and arrow, cannon, necessitating protection gear such as shields or body armour (cf. Diamond, 1997). Another technique is the effort to enhance the biological capabilities of soldiers, a notorious example of which is the use of the methamphetamine Pervitin in World War II (Rasmussen, 2011). The foundational ideas of optimising military strategy (Von Clausewitz, 1942) are instantiated in cultural techniques such as war-gaming, academic approaches to capture the dynamics of war empirically (Bousquet, Grove & Shah, 2020) or the computer-assisted simulation and prediction of military scenarios today (Cayirci et al, 2022).

In both historical narratives, technology entertains notions of power and (loss of) control, either taming nature or subjugating enemies by enhancing the soldier and its abilities. Technology represents both magical, sublime or social qualities (Appadurai, 1986) or can elicit horror or repulsion, running the risk of rendering the human obsolete, a destruction even beyond imagination (Anders, 2002). It is these histories in which the cultural portrayal of autonomous weapons is rooted and finds its expression. For example, science-fiction films and public campaigns cater to doomsday scenarios that mobilise pictures of merciless and destructive machines. AWS are pictured as "killer robots" (Stop Killer Robots, n.d.) or "Slaughterbots" (Autonomous Weapons, n.d.). The idea of AI "going rogue", turning against its makers and humanity at large, is another common trope of the theme of loss of control and taming. Autonomous and human-like machines evoke fears of a lethal intelligence that outsmarts humans. The (real) opacity of these AI-based systems, which cannot be comprehended by the majority of people fosters the idea of networked architectures making themselves independent and take a "life" (hence, becoming wild nature) of their own. This combination of the incomprehensible and an unvisited intelligence evokes strong sentiments

of both fascination and horror. Certainly, a great deal of the intimidation evoked by the sublime aura of AWS is produced through the limitless force of human imagination, quickly crossing the boundaries of fact and fiction. Take motifs of a sinister "HAL 9000" computer in *Space Odyssey*, the idea of a cybernetic android killer such as "the Terminator", or scenarios of killer drone swarms (also depicted in the video "Slaughterbots", see above), which reverberates with Alfred Hitchcock's menacing motif in "*The Birds*"). These portrayals of fictional destructive lethal machinery are sustainably shaping the public and political perceptions of AWS and are contributing to a large extent to their popularity.

3. Imaginations of autonomous weapons are utilised as tools and rhetorical devices of geopolitical aspirations – and provide a smokescreen for other fields of conflict and warfare

Putting into perspective the current detrimental effects of AWS, it is certainly noteworthy that conventional firearms – at the time of writing in 2024 – inflict more harm and human suffering than AI-assisted military technologies. In the US alone the latest complete data shows that in 2021, 48,830 people died from gun-related violence (Gramlich, 2023). In Mexico, official numbers declare 22,309 gun related deaths in 2022 (Álvarez, 2023), and in South Africa, 8,388 deaths in 2021, with numbers on the rise, as alone between October and December 2022, more than 7,500 people died through firearms (Kirsten, 2023 & Khumalo, 2023). In 2022, in the US alone the firearm and ammunition industry was responsible for as much as $80.73 billion in total economic activity of the country (NSSF, 2022). In comparison, in the same year, the *global* military artificial intelligence market size was substantially smaller, valued at $7.4 billion in 2022 (Grand View Research, n.d.). Pistols and rifles seem to be perceived as conventional, almost traditional, and are more accepted among the public. They have been widely disseminated and decentralised in use around the globe for decades, are comparably low-tech engineered and remain largely unchecked in trade – despite a global Arms Trade Treaty, which has not been signed by nations with major production sites (Amnesty International, n.d.). There seems to be a surprising disconnect between the highly differentiated debates on future warfare, the subsequent risk scenarios and elaborated assessments of ethical repercussions and the needs for political and legal regulation – and the attention devoted to the risks and harms of contemporary conventional weapons. While the

latter are far from being accepted, conventional weapons are discussed alongside very conventional arguments. They lack the nimbus of glitzy AI-enabled future warfare.

The main difference seems to be that the rhetorical drumbeat around autonomous weapons is already part of modern warfare and an effective tool in political communication. Suggestions of AI capabilities, woven into the political rhetoric of state actors can be an effective vehicle in strategic deterrence of enemies (Johnson, 2020). The praise of AWS capabilities can therefore be understood as a means of psychological warfare, with the aim to clarify one's position in the geopolitical order and strategically contain, defend or strive for hegemonic aspirations. As argued elsewhere, the comparison between Chinese and US AWS imaginaries shows that "[military] AI is in both cases regarded as a means to realise these socio-political ideals, with supremacy achieved by technological prowess being a shared theme for both" (Bächle & Bareis, 2022: 7).

At the time of writing, recent examples of attempts to foreground a branding of AI use in military contexts can be found in the employment of target recommender systems. The Israel Defense Forces (IDF) use AI in the military operations in Gaza following the terrorist attacks by Hamas on Israeli civilians on October 7, 2023. The employed AI system is called "Habsora", the "gospel" platform – a "holy message" in biblical terms. It is this recommender system used for enemy detection, which plays "a critical role in building lists of individuals authorised to be assassinated" by airstrikes (Davies et al, 2023). Ukrainian forces use recommender systems, so called Geographic Information System Art for Artillery (GIS ARTA, n.d.) for fire missions, also being coined by its Ukrainian developer Sherstyuk "Uber for Artillery" (Bruno, 2022). GIS Arta speeds up artillery missions by sourcing real-time data "from drones, targets reported by forward observers armed with cell phones, counter battery radars, and satellite-based imagery" (Zikusoka, 2023, n.p.).

The references to different motifs and imaginaries are meant to reach objectives in political communication – but as a side effect complicate understandings of military AI and AWS in public, academic or political spheres. Their meanings get loaded with associations borrowed from religious or fictional texts. The technology is subjected to interpretations in an arena that is already characterised by strivings for dominance in public discourse. As another side-

effect, the overemphasis on the imagined potentials of modern intelligent weapons shifts the focus away from the very conventional and often very 'stupid' weapons – such as mass-produced simplistic drones (for example the Iranian-Russian cooperation to produce Shahed-136 drones to attack Ukraine; Bennett & Ilyushina, 2023) – which pose a threat by way of sheer quantity and easy access, as they can also be manufactured or commissioned by non-state actors.

Despite all this, AWS are by no means limited to a rhetorical realm but also play a considerable role in warfare, reinforcing and executing state interests. Private armies or military contractors – so-called irregular militaries – are characteristic of neoliberal modes of warfare. Easy-access and high quantity automatic weapons must be regarded as a particular threat in the hands of these non-state actors, employing harmful technologies outside of regulatory frameworks. AWS, being software-based to a large extent, makes the dissemination of harmful technology easier (often in a downloadable, intangible form) and at the same time more difficult to trace compared to conventional weapons.

From the view of international relations, AWS can be seen as a continuation of a prerogative of state violence that transcends national borders, and acts as an event outside of temporal and spatial limitations. For example, Rooke argues in this volume that the US-Air Force's "air-mindedness" executes state violence in a "hierarchical ordering that places the US at the top of this dominant spatiotemporality". From this perspective, AWS in the form of drones and other unmanned warfare like cyber attacks resemble a form of warfare that executes power through writing and simplistic categorising (enemy/ally; hostage/terrorist). It is the power to make (dis)appear perpetrators, victims, violence, sufferings, injustices, as they happen far away from the auspices of international humanitarian law, human rights and public accountability. Rupka and Baggiarini argue that air warfare conducted through drones resembles a "militarised gaze (...) [which] is both everywhere and nowhere, whilst its power successfully enables the rendering of "populations into the terrain of state legibility and security so that they might become governable subjects." (Rupka & Baggiarini, 2018: 13). Without an official declaration of war, hegemonic states can operate effectively in the geopolitical realm without holding accountability for their actions. Violence acts without

having troops on the ground and outside a normative international system. This converts drone and AWS violence by states into a clandestine non-event.

Regarding their symbolism, rhetorics and kinetic abilities, AWS are useful for various geopolitical aspirations of states. Thereby they also provide a smokescreen for other fields of, often, more conventional conflict and warfare around the globe.

4. Autonomy in weapon systems emphasises the necessity to thoroughly theorise AI

The ongoing efforts to regulate autonomous weapons and the use of artificial intelligence has not just underlined the need to properly define what makes an autonomous weapon system really *autonomous* or what is characteristic of an AI system that sets it apart from its technological precursors. In a more abstract sense, it also puts a spotlight on the many, still remaining conceptual voids surrounding current debates on autonomous systems and AI.

The rise of AI, especially accelerated by a combination of machine learning (ML) data processing capabilities, more effective sensors and advanced infrastructure, weapon systems are able to operate with much less human intervention than the preceding technologies could. The allure of AI has seemingly changed attributes from *automatic* into *autonomous* systems, which sparks epistemic but also regulatory confusion (Sauer, 2016). From a disciplinary standpoint, autonomy has always been a contested concept. Also in technical and engineering discourses it has become a widely used term, where it commonly evokes associations of independence, intelligence, self-governance, self-sufficiency, the ability to learn and adapt (e.g. orientation in unknown, unstructured and dynamic environments) or the execution of self-determined decisions (Williams, 2015). Such functional viewpoints in engineering, easily conflate understandings of autonomy, trust and responsibility from the viewpoint of human moral agency (see Schwarz in this volume). As a consequence, and problematically so, technical understandings are starting to be applied in the realm of human ethics, resulting in a mechanical weighing of human value similar to mathematical calculation and algorithmic optimisation.

Besides the conceptual vagueness, the terminology applied in the discussions on AWS are commonly contextualised in larger cultural narratives. Here, notions of machine autonomy in weapons are contested and often embedded in fictional narratives. They utilise broadly-known mythological and anthropomorphic allusions or borrow motifs from popular culture. For example, the US counter rocket, artillery and mortar (C-RAM) close-in weapon system "Phalanx", in service since the 1980s, takes its reference from the ancient Greek empire, where spears units formed a phalanx formation in battle against the enemy. The C-RAM vulcan cannon can be mounted on ships, and, next to the Greek reference for its name, Navy's crews gave the Phalanx systems the pet name "R2-D2" because their appearance is reminiscent of the droid R2-D2 from the Star Wars films (Stoner, 2009).

It seems common practice among military and political stakeholders to re-interpret the concept of autonomy and AI to particular means, which often comes at the cost of nullifying the conceptual or practical use of the term. A position paper submitted in 2018 to the CCW negotiations in Geneva by the German delegation, for example, states the following: "Having the ability to learn and develop self-awareness constitutes an indispensable attribute to be used to define individual functions or weapon systems as autonomous" (Permanent Representation of the Federal Republic of Germany to the Conference on Disarmament in Geneva, 2018). Tying "self-awareness" to a definition of machine autonomy is absurd, for obvious reasons. It can, however, have a rhetorical function at the negotiation table. In the same year, the Chinese delegation at the CCW defined a necessary feature of AWS with the following condition: "once started there is no way to terminate the device" (CCW Group of Governmental Experts on LAWS, 2018: 1). This entertains the no less absurd scenario of an AI gone rogue, completely outside of human control. Partly due to the terminological confusion and strategic vagueness, the CCW negotiations have been gridlocked, far from reaching a consensus that would honour International Humanitarian Law in a serious attempt to regulate the actual reality of autonomous weapon systems (see also Suchman in this volume). Overall, some public and military interpretations of autonomy in the AWS debate articulate sensationalist fiction and have succeeded in capturing not only public discourses (see Cave & Dihal, 2019; Campolo & Crawford, 2020), even debates in research (Natale & Ballatore, 2020), and have also found their way in the regulatory arena (see Bächle & Bareis, 2022.)

Also the more conceptually grounded notions of autonomy in automated warfare are no historically fixed constants but subject to change. Ernst, for example, argues in this volume that rather than dealing with self-sufficient and autonomous battle machines such as drones, tanks or ships, autonomy in contemporary military visions is better understood as resilient *networks* between connected agents and infrastructures. "Combat clouds" engage in warfighting, highlighting the importance of communication hubs or real-time data analytics. Projects such as the European Future Combat Air System (FCAS) also point in this direction (see Hälterlein in this volume). These examples of recommender systems or combat clouds highlight the various elements in warfare that are increasingly automated, hence, different from the idea of a self-sustained "autonomous" battle machine. Procedures of identifying, selecting and determining who is a civilian or an enemy (see Packer & Reeves in this volume) or practices of tracking or engaging with targets are being automated through algorithms.

Contrasting with many of the prevalent approaches used in political science, law or philosophy, which understand autonomy as a distinct quality associated with the human condition, these examples also indicate that autonomy rather emerges performatively within social or material structures and is thus subject to cultural change and national differences (Haraway, 2006). A performative understanding of autonomy also helps to look past many of the thought experiments that consider a world, in which humans will finally have acquired human-like abilities. It sheds light on the *mechanisms* that provoke what could be called "autonomy effects": Databases in which target lists are stored, tracking and target selection mechanisms, computer programmes that control when systems should no longer listen to human actions and so forth. It is not only important to unpack the metaphorical uses and the practices of how autonomy is "made" (Noorman & Johnson, 2014) but also makes visible the networked and automated infrastructures that underlie imaginations around LAWS.

It is exactly this interpretative openness of the term autonomy that predestines it to be applied in various contexts and with tailor-made meanings. The erosion of its semantic qualities not just calls for a thorough reflection of the premises used but even more importantly for a theorisation of AI in general.

5. Autonomous weapons challenge our understanding of what is human and foreground the relationship between humans and machines

As part of the shift away from solely looking at the suggested autonomy of a distinct system and in favour of taking into account the performative dimension and underlying structures of autonomy, it is particularly necessary to assess the human/machine relationship. Conceptually, reality of the existing "human-machine autonomies" (Suchman & Weber, 2016) – rather than autonomous machines – have important roots in cybernetic theory, establishing an analogy between humans and machines via a universally applicable analogy: "The systems analogy, as well as the understanding of systems as goaldirected and purposeful, is a central precondition for the idea of the 'autonomy' of so-called smart and intelligent (war) machines." (Suchman & Weber, 2016, 83-4).

While the human/machine systems analogy is a theoretical precondition of common ideas of autonomy and autonomous weapons – often drawing false equivalencies, as discussed in the previous section – it paradoxically also elicits the paradigmatic question on differentiating humans and machines. In the most basic terms, this means asking about the human element, whether it being part of "the loop" or in "meaningful control". Imagining weapons necessarily entails imagining a version of the human, their role in the relation with machines, as in ethical, political or legal categories: when and how should a human be able to intervene, should a human necessarily be involved in the decision to kill another human, and so forth?

On par with this, the military discourse on autonomous weapons is no longer purely techno-centric but moves towards both the human/machine relationship or even human centricity. "Manned/unmanned teaming", "human augmentation" (UK Ministry of Defence/Bundeswehr Office for Defence Planning, 2021) or "the enhanced soldier" (de Boisboissel & Le Masson, 2021) both take into account and shape this technological, conceptual and strategic shifts. Augmentation has even been identified as the up-and-coming paradigm in discussions of autonomous weapons and military AI (cf. Favaro & Schwarz, 2022).

"The human" has always been present in a functional sense, because it is a vital – but often only pro forma – point of reference. Debates on political, legal or ethical debates on

responsibility, dignity, intentionality, etc. require a human to pin them on: As long as "the human" as a function is formally in the picture, the otherwise autonomous machine seems more legitimate.

It is high time, however, to direct our attention to humans, which means rather than solely discussing autonomous weapons as technical entities, we need to focus on human/machine interactions and relations while fully acknowledging that fully autonomous systems are, even though they foster our fascination and horror, a rather skewed narrative.

The book's sections and individual contributions

The book's structure introduces three individual sections that engage with current realities of autonomous weapons. Each section analyses autonomous weapons from a particular trope of perspective: 1. Fictions, Narratives & Theories, 2. Technologies & Materialities, and 3. Politics & Ethics. The beginning of each section is introduced by an artist and their vision on autonomous weapons. The sectioning adheres to an analysis of the different meanings articulated across these domains that constitute the realities of AWS and powerfully influence how we perceive and engage with this technology.

Section 1. Narratives & Theories

This section looks at cultural texts that are marked as fiction (e.g. science-fiction films and novels etc.) as well as those marked as non-fiction in research. Its goal is to analyse the potentials, risks, narratives and aesthetics that are associated with AWS:

- *ARTWORK*. «The Unreachable Myth»**,** Killing unknown victims, with unsensible ways by unidentified perpetrators for unapparent reason. By Yinyu Wang, 2023
- Jennifer Rooke: *The AI-Lure of US Airpower: Imaginaries of Disruption in the Pursuit of Technological Superiority Since the Early 20th Century*
- Rebecca Jones: *From Maschinenmensch to Robot Bubs: Female-Presenting Autonomous Weapons Systems in Live-Action Films from 1927-2022*

- Teresa Heffernan: *Autonomous weapons in fiction and the fiction of autonomous weapons*

- Ingvild Bode & Shimona Mohan: *From the Reel to the Real: Narratives of Weaponised Artificial Intelligence Technologies in India*

In *"The AI-Lure of US Airpower: Imaginaries of Disruption in the Pursuit of Technological Superiority Since the Early 20th Century",* Jennifer Rooke analyses the military imaginaries that shape the use of automated pattern and target recognition technologies by the US Air Force within their intelligence, surveillance and reconnaissance operations. She traces how the US air-mindedness emerged and developed into a hegemonic prerogative to achieve air superiority by political, legal and technical means around the world. The article by Rebecca Jones, *"From Maschinenmensch to Robot Bubs: Female-Presenting Autonomous Weapons Systems in Live-Action Films from 1927-2022"* looks at the evolution of AWS through cinematic history with a particular focus on female representations of weapons in humanoid form. While weapons are commonly associated with male representations (with the Terminator as the most common trope), the representation of warfare is highly gendered. "Female-presenting autonomous weapons" mirror the patriarchal gazes of their times that are merged with technical features that saliently negotiate stereotypical imaginations of the female. Jones analyses how female-presenting AWS negotiate fears and hopes of subordination, domination, or (loss of) control, once more stressing how gender, power and the technical are constantly reworked with AWS. Teresa Heffernan's analysis *"Autonomous weapons in fiction and the fiction of autonomous weapons"* also looks at the domain of fiction. She poses the question how the literal readings of fiction to animate real machines distracts from the real-world development of this technology. By taking reference to Karel Çapek's play *R.U.R.: Rossum's Universal Robots* (1920) and James Cameron's *The Terminator* (1984) and its sequels, she shows how fiction has long connected the fetishisation of this technology to for-profit research and development. Ingvild Bode and Shimona Mohan take the reader to a completely different geographical part of the world and interrogate in "*From the Reel to the Real: Narratives of Weaponised Artificial Intelligence Technologies in India"* public perspectives on AWS. Analysing survey data collected in January 2023, they find that weaponised AI narratives of Anglophone countries have a high resonance among Indian

respondents. At the same time, Indian respondents also shared distinct ways of narrating AI technologies that integrate cultural particularities, drawing for example on Indian mythology and folklore as well as the mixing of genres that are typical of most Indian movie productions.

2. Technologies & Materialities

This section looks at the concepts that are frequently applied when explaining the technological and material particularities of AWS. These include specific notions of decision-making, technological agency or autonomy and debates around human-machine entanglements such as 'meaningful human control'. At the same time, the discourses on weapons technologies are always historically interwoven with the conceptual transformation of warfare and show how materialities influence particular military doctrines and vice versa:

- *ARTWORK.* «Transformator». By Peter Behrbohm, since 2013
- Lucy Suchmann: *Il/legal war: Expanding the frame of meaningful human control from military operations to democratic governance*
- Christoph Ernst: *From network-centric warfare to autonomous warfighting networks – Recontextualising AWS imaginaries*
- Jens Hälterlein: *Governing autonomies – Imagining responsible AI in the European armament project "Future Combat Air System"*
- Jeremy Packer & Joshua Reeves: *New media, new enemies: The emergence of automated weapons in counterterrorism*

In "*Il/legal war: Expanding the frame of meaningful human control from military operations to democratic governance",* Lucy Suchman comments on the viewpoints on the legality of AWS. She scrutinises the debates of war that sustain militarism and how they might be challenged, not only from within but also beyond the project of arms control. Suchman draws from her own 2016 testimony at the UN Convention on Certain Conventional Weapons (CCW) where she argued against the capacity of AWS to adhere to International Humanitarian Law. In her article she puts forwards requirements of situational awareness and adherence to the principle of distinction as a necessary condition for lawful autonomy that remain unfulfilled

by AWS. Christoph Ernst also points to the complicated picture of autonomy and human-machine entanglement in *"From network-centric warfare to autonomous warfighting networks – Recontextualising AWS imaginaries".* He argues that the relevance of network-centricity for AWS imaginaries and the associated visions of future warfare is often overlooked. Ernst shows how ideas on network-centric warfare developed during the 1990s and early 2000s are the historical origins, which provide important scripts and metaphors for contemporary AWS debates. By tracing this historical legacy he argues that current AWS imaginaries contain the infrastructural vision of what can be called "autonomous warfighting networks". Jens Hälterlein applies these notions of networked warfare to a concrete case study in Europe. In *"Governing autonomies – Imagining responsible AI in the European armament project "Future Combat Air System (FCAS)",* he analyses how the FCAS project imagines AI in the year 2040 as the means to enhance human decision-making and to enable responsibility and accountability. By scrutinising the so-called FCAS Ethical AI Demonstrator, he shows how FCAS applies a liberal anthropology, featuring individual responsibilisation of operators and environmental management of behaviour through ethics by design – which, in his view, fails to live up to FCAS' own claims of enhancing human responsibility and accountability. The section concludes with "*New Media, New Enemies: The Emergence of Automated Weapons in Counterterrorism",* in which Jeremy Packer and Joshua Reeves dive into the recursive relationship between media technology, knowledge creation and the production of political and military enemies. Through the prism of media theory they show how media technologies produce new ways of perceiving the surrounding world and the threats that lurk therein. When applied in political or military contexts, they argue, it means that enemies will always be uncovered, as enhanced visibility automatically brings new enemyship to the surface. They observe that with positive feedback systems, there is no way to ultimately find and neutralise all enemies. The system's operation demands the constant discovery of new problems to solve.

3. Politics & Ethics

This section looks at the understandings and meanings of LAWS that are applied in political and ethical contexts, which are often based on 'as if' scenarios. Translated into political action,

these meanings and their underlying assumptions create realities in their own right. While the actual technological capabilities are still limited, their anticipated futures have nonetheless severe implications for global security policies, regulatory and legal initiatives or military operations in light of their use by states as well as non-state actors.

- *ARTWORK*. «XCI|XCIX, (91|99)». By Johannes Weilandt, 2023.
- Elke Schwarz: *Engineering moral failure? The challenges of algorithmic ethics for lethal autonomous weapon systems*
- Bernhard Seidl: *Legitimising and contesting Lethal Autonomous Weapons Systems in Japan: A multi-layered analysis of public discourse*
- Jutta Weber: *The reality of (past) Future Air Combat Systems. On climate wars, carbon costs and rare earth elements*
- Thomas Christian Bächle & Jing Zeng & N.N.: *Autonomous weapons discourses in Chinese state media*

In her contribution "*Engineering moral failure? The challenges of algorithmic ethics for lethal autonomous weapons",* Elke Schwarz observes that over a decade's worth of discussions on the ethical and legal implications of autonomous weapons systems have yielded limited results. Problematically, these discussions are marred by unhelpful conflations, with both human agency and machine agency being read through a technological lens wherein functional equivalences are drawn between the two. She examines these discourses and their logical foundations and argues that rather than helping make sense of the specific demand of moral agency and responsibility in the context of AWS, they take us further away from understanding moral concerns as exclusively related to humans. The political and ethical understanding of AWS remains contested. Not only ethically, as Schwarz shows, but also from the viewpoint of political institutions across the globe. Bernhard Seidl conducts an analysis of public discourse on lethal autonomous weapon systems (LAWS) in Japan. He examines texts produced in or for the public sphere, including policy documents, NGO material and newspapers, in order to understand how the adoption of LAWS in Japan is legitimised and

contested. With *"Legitimising and contesting Lethal Autonomous Weapons Systems in Japan: A multi-layered analysis of public discourse",* Seidl places his findings in the context of the nation's evolving security identity and reveals the interplay between the discourse layers and actors, realised in a language influenced by facts and imaginaries particular to the Japanese context. Evoking so much attention and allure, AWS not only have the power to attract political state interest – they also mute and sideline their hazardous side-effects. In *"The reality of (past) Future Air Combat Systems. On climate wars, carbon costs and rare earth elements",* Jutta Weber discusses the carbon costs, greenhouse gas emissions and the rare earth metal dependencies of present and future military systems. She emphasises that world's militaries and associated military technology industries are responsible for around 5.5 percent of global greenhouse gas emissions (GHG) – without counting post-war recovery. Looking concretely at Future Combat Air programmes and the realities of their development and deployment in the future, she argues that the emissions will ultimately inhibit the realisation of these systems, rendering their future something that has already passed.

Filmography

2001: A Space Odyssey (1968) Directed by Stanley Kubrick, USA: Metro-Goldwyn-Mayer.

The Birds (1963) Directed by Alfred Hitchcock, USA: Alfred Hitchcock Productions.

The Terminator (1984) Directed by James Cameron, USA: Hemdale, Pacific Western Productions, Euro Film Funding & Cinema '84.

References

Álvarez, G. (2023) 'Armas de fuego en México: pensar la violencia a partir de sus medios de ejercicio', *Este Pais*, [online] 4 December, Available from: https://estepais.com/tendencias_y_opiniones/armas-fuego-mexico-pensando-violencia-sus-medios-ejercicio/ [Accessed 12 March 2024].

Amnesty International (n.d.) 'Arms Control', *Amnesty International*, [online], Available from: https://www.amnesty.org/en/what-we-do/arms-control/ [Accessed 11 March 2024].

Anders, G. (2002) *Die Antiquiertheit des Menschen: über die Zerstörung des Lebens im Zeitalter der dritten industriellen Revolution* (2nd edition), München: CH Beck.

Antinori, A. (2019) 'Terrorism and DeepFake: from Hybrid Warfare to Post-Truth Warfare in a Hybrid World', in P. G Griffiths & M. N. Kabir (eds) *Proceedings of the European Conference on Impact of Artificial Intelligence and Robotics,* Oxford, U.K.: EM-Normandie Business School.

Appadurai, A. (1986) The Social Life of Things. Commodities in cultural perspective, Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819582

Arkin, R. C. (2009) *Governing lethal behaviour in autonomous robots*, New York: CRC Publishers.

Arquilla, J. (2021) *Bitskrieg: The New Challenge of Cyberwarfare*, Cambridge, UK: Polity Press.

Arquilla, J and Ronfeldt, D. (2000) *Swarming and the future of conflict*, Santa Monica, CA: RAND Corporation & National Defense Research Institute.

Autonomous Weapons (n.d.) 'Slaugherbots are here.', *Autonomous Weapons* [online], Available from: https://autonomousweapons.org/ [Accessed 05 March 2024].

Barber Associates (1975) 'The Advanced Research Projects Agency, 1958-1974. II-10', in: Defense Technical Information Center. Available from: https://archive.org/details/ARPAhistory/page/n5/mode/2up [Accessed 12 March 2024].

Bächle, T.C. (2023) 'The age of machine autonomy?', *Digital society blog*. https://doi.org/10.5281/zenodo.8273068

Bächle, T.C., Bareis, J. (2022) '"Autonomous weapons" as a geopolitical signifier in a national power play: analysing AI imaginaries in Chinese and US military policies', *European Journal of Futures Research* 10(20): 1-18. https://doi.org/10.1186/s40309-022-00202-w

Bareis, J., & Katzenbach, C. (2022) 'Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics', *Science, Technology, & Human Values*, 47(5): 855-81. https://doi.org/10.1177/01622439211030007

Bareis, J., Roßmann, M., & Bordignon, F. (2023) 'Technology hype: Dealing with bold expectations and overpromising', *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 32(3): 10-71. https://doi.org/10.14512/tatup.32.3.10

Baudrillard, J. (1995) *The Gulf War Did Not Take Place*, Bloomington and Indianapolis: Indiana University Press.

Bennett, D., Ilyushina (2023) 'Inside the Russian effort to build 6,000 attack drones with Iran's help', *The Washington Post*, [online] 17 August, Available from: https://www.washingtonpost.com/investigations/2023/08/17/russia-iran-drone-shah

ed-alabuga/ [Accessed 11 March 2024].
Beckert, J. (2016) *Imagined Futures*, Cambridge, MA: Harvard University Press.
Bhuta, N., Beck, S., Geiß, R., Liu H.Y. and Kreß, C. (2016) Autonomous Weapons Systems: Law, Ethics, Policy, Cambridge: Cambridge University Press.
Block, F. (2008) 'Swimming Against the Current: The Rise of a Hidden Developmental State in the United States', Politics & Society, 36(2): 169-206. https://doi.org/10.1177/0032329208318731
Bousquet, A., Grove, J., & Shah, N. (2020) 'Becoming war: Towards a martial empiricism', Security Dialogue, 51(2-3): 99-118. https://doi.org/10.1177/0967010619895660
Bruno--, M. (2022) '"Uber For Artillery" – What is Ukraine's GIS Arta System?', The Moloch, [online] 24 August, Available from: https://themoloch.com/conflict/uber-for-artillery-what-is-ukraines-gis-arta-system/ [Accessed 11 March 2024].
Campolo, A., Crawford, K. (2020) 'Enchanted Determinism: Power without Responsibility in Artifical Intelligence, Engaging Science, Technology, and Society 6: 1-19. https://doi.org/10.17351/ests2020.277
Cave, S., Dihal, K. (2019) 'Hopes and fears for intelligent machines in fiction and reality', Nature machine intelligence, 1(2): 74-8 https://doi.org/10.1038/s42256-019-0020-9
Cayirci, E., AlNaimi, R. and AlNabet, S. S. (2022) 'Computer Assisted Military Experimentations', in 2022 Winter Simulation Conference (WSC), Singapore, pp 1311-324. http://dx.doi.org/10.1109/wsc57314.2022.10015294
CCW Group of Governmental Experts on LAWS (2018) 'Position Paper', Reaching Critical Will, [online] April, Available from: https://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.1-WP7.pdf [Accessed 07 March 2024].
Cebrowski, A.K. (2005) 'The Implementation of Network-Centric Warfare', Office of Force Transformation Washington, [online], Available from: https://apps.dtic.mil/sti/citations/ADA446831 [Accessed 07 March 2024].
Coeckelbergh, M. (2011) 'From killer machines to doctrines and swarms, or why ethics of military robotics is not (necessarily) about robots', Philosophy & Technology 24(3): 269-78. https://doi.org/10.1007/s13347-011-0019-6
Cohen, E. A. (1996) 'A Revolution in Warfare', Foreign Affairs 75(2): 37-54. https://doi.org/10.2307/20047487.
Cummings, M. L. (2017): „Artificial Intelligence and the Future of Warfare". URL: https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf, last visited 2020-20-01.
DARPA (n.d.) 'Homepage', Darpa, [online], Available from: https://www.darpa.mil/ [Accessed 12 March 2024].
DARPAtv (2018) 'Darpa Overview', *YouTube*, [Video] 16 January, Available from: https://www.youtube.com/watch?v=IcV4pgB5QY0 [Accessed 12 March 2024].
Davies, H., McKernan, B. & Sabbagh, D. (2023) ''The Gospel': how Israel uses AI to select bombing targets in Gaza', *The Guardian*, [online] 1 December, Available from: https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-aito-select-bombing-targets [Accessed 05 March 2024].
De Boisboissel, G., & Le Masson, J.-M. (2021) 'The Enhanced Soldier. Definitions', *Military Review*: 1-5. Available from: https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2021F

rench-OLE/Part-2-Definitions/ [Accessed 12 March 2024].
De Landa, M. (1991) War in the Age of Intelligent Machines, New York, NY: Zone Books.
Diamond, J. M. (1997) Guns, Germs, And Steel: The Fates of Human Societies, New York: W. W. Norton.
Eaves, W. (2018) Murmur, London: CB Editions.
Echevarria, A. J., Shaw, J. M. (1992) 'The New Military Revolution: Post-Industrial Change', Parameters Winter 22(1): 70-79. https://doi.org/10.55540/0031-1723.1624
Elish, M.C. (2018) 24/7: Drone Operations and the Distributed Work of War, New York, NY: Columbia University in the City of New York.
Ernst, J., Sasse, B. (2016) 'Midlands Voices: World needs United States to lead', Omaha World-Herald, [online] 16 February, Available from: https://omaha.com/opinion/midlands-voices-world-needs-united-states-to-lead/article_74c0f92a-04a1-5ff1-9314-ff24a3030537.html[Accessed 07 March 2024].
Favaro, M., & Schwarz, E. (2022) 'Human Augmentation and Nuclear Risk: The Value of a Few Seconds*, Arms Control Today 54.
Galison, P. (1994) 'The Ontology of the Enemy: Norbert Wiener and the Cybergenetic Vision', Critical Inquiry, 21(1): 228-66.
GIS ARTA (n.d.) 'GIS "ARTA": automated command and control system', GIS ARTA, [online], Available from: https://gisarta.org/en/index.html [Accessed 05 March 2024].
Goecks, V. G., Waytowich, N. (2024) COA-GPT: Generative Pre-trained Transformers for Accelerated Course of Action Development in Military Operations, NATO Science and Technology Organization Symposium (ICMCIS) by the Information Systems Technology (IST) Panel, Koblenz: Germany. https://doi.org/10.48550/arXiv.2402.01786
Graae, A.I., Maurer, K. (2021) Drone Imaginaries: The Power of Remote Vision, Manchester: Manchester University Press.
Gramlich, A. (2023) 'What the data says about gun deaths in the U.S.', Pew Research Center, [online] 26 April, Available from: https://www.pewresearch.org/short-reads/2023/04/26/what-the-data-says-about-gun-deaths-in-the-u-s/ [Accessed 11 March 2024].
Grand View Research (n.d.) 'Artificial Intelligence In Military Market Size, Share & Trends Analysis Report By Offering, By Application, By Technology, By Platform, By Installation, By Region, And Segment Forecasts, 2023 - 2030', Grand View Research, [online], Available from: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-military-market-report# [Accessed 12 March 2024].
Guardian Alliance Technologies (2021) 'How a retired cop is helping to solve the law enforcement hiring crisis through technology', Cision PR Newswire, [online] 26 March, Available from: https://www.prnewswire.com/news-releases/how-a-retired-cop-is-helping-to-solvethe-law-enforcement-hiring-crisis-through-technology-301256468.html [Accessed 07 March 2024].
Haraway, D. (2006) 'A Cyborg Manifesto', in S. Stryker and S. Whittle (eds) The Transgender Studies Reader, New York: Routledge, pp 103-18. https://doi.org/10.4324/9780203955055
Horwitz, R. B. (2022) 'Trump and the "deep state"', in T. S. James (ed) The Trump Administration, London: Routledge, pp 41-58. https://doi.org/10.4324/9781003259923
Hoskins, A., O'Loughlin, B. (2015) 'Arrested war: The third phase of mediatization',

Information, Communication and Society 18(11): 1320-338. https://doi.org/10.1080/1369118X.2015.1068350
Hughes, J. (2004) 'Deconstructing the bomb: recent perspectives on nuclear history', The British Journal for the History of Science, 37(4): 455-64. https://doi.org/10.1017/S0007087404006168
Jasanoff, S. (2015) 'Future Imperfect: Science, Technology, and the Imaginations of Modernity, in S. Jasanoff & S.-H. Kim (eds) Dreamscapes of Modernity, Chicago and London: University of Chicago Press, pp 1-33. https://doi.org/10.7208/9780226276663
Jasanoff, S., Kim, S.-H. (2009) 'Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea', Minerva 47: 119–46. https://doi.org/10.1007/s11024-009-9124-4
Jasanoff, S., Kim, S.H. (2015) Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power, Chicago and London: University of Chicago Press.
JEDI Foundation (n.d.) 'Homepage', Joint European Disruptive Initiative, [online], Available from: https://www.jedi.foundation/ [Accessed 05 March 2024].
Johnson, J. (2020) 'Deterrence in the age of artificial intelligence & autonomy: a paradigm shift in nuclear deterrence theory and practice?', Defense and Security Analysis 36(4): 422–48. https://doi.org/10.1080/14751798.2020.1857911
Karppi, T., Böhlen, M. and Granata, Y. (2018) 'Killer robots as cultural techniques', International Journal of Cultural Studies 21(2): 107–23. https://doi.org/10.1177/1367877916671425
Khumalo, T. (2023) 'South Africa: Spike in gun crime angers citizens', Deutsche Welle, [online] 7 March, Available from: https://www.dw.com/en/south-africa-spike-in-gun-crime-angers-citizens/a-64903654[Accessed 12 March 2024].
Kirsten, A. (2023) 'Guns and gun violence in South Africa — the evidence', Daily Maverick, [online] 16 February, Available from: https://www.dailymaverick.co.za/article/2023-02-16-guns-and-gun-violence-insouth-africa-the-evidence/[Accessed 12 March 2024].
Krasmann, S., Weber, J. (2015) 'Game changer? On the epistemology, ontology and politics of drones', Behemoth A Journal on Civilisation 8(2): 3-11- https://doi.org/10.6094/behemoth.2015.8.2.866
Krishnan, A. (2016) Killer robots: Legality and ethicality of autonomous weapons, London and New York: Routledge
Latour, B. (1993) We have never been modern, Cambridge, Massachusetts: Harvard University Press.
Lenoir, T., Caldwell, L. (2018) The Military-Entertainment Complex. Cambridge, MA and London: Harvard University Press.
Lin, P., Bekey, G. and Abney, K. (2008) 'Autonomous Military Robotics: Risk, Ethics, and Design'. Cal Poly, [online] 20 December, Available from: https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac [Accessed 07 March 2024].
Lubell, N. (2023) 'Keynote Lecture: AI in Critical Decision: is Human Judgment a Legal Requirement?', DILEMA 2023 Conference, [lecture], 12 Octobre, Program available from: https://www.asser.nl/media/796679/dilema-2023-conference-programme.pdf [Accessed 13 March 2024].
Lyon, D. (2015) Surveillance after Snowden, Hoboken: John Wiley & Sons.

Marx, L. (2000) The Machine in the Garden: Technology and the Pastoral Ideal in America, New York: Oxford University Press.
Mazzucato, M. (2011) 'The entrepreneurial state', Soundings 49(49): 131-42. https://doi.org/10.3898/136266211798411183
Merrin, W. (2018) Digital War: A Critical Introduction, London and New York: Routledge.
Natale, S., & Ballatore, A. (2020) 'Imagining the thinking machine: Technological myths and the rise of artificial intelligence', Convergence, 26(1), 3-18. https://doi.org/10.1177/1354856517715164
Noorman, M., Johnson, D.G. (2014) 'Negotiating autonomy and responsibility in military robots', Ethics Inf Technol, 16: 51–62. https://doi.org/10.1007/s10676-013-9335-0
NSSF (2022) 'Firearm and Ammunition Industry. Economic Impact', NSSF, [Report], Available from: https://www.nssf.org/wp-content/uploads/2022/03/2022-Firearm-Ammunition-Industry-Economic-Impact.pdf [Accessed 12 March 2024].
Nye, D. E. (2004) 'Technological Prediction: A Promethean Problem', in M. Sturken, D. Thomas and S. Ball-Rokeach (eds) Technological Visions: The Hopes and Fears That Shape New Technologies, Philadelphia, PA: Temple University Press, pp 159-76.
Permanent Representation of the Federal Republic of Germany to the Conference on Disarmament in Geneva (2018) 'Statement delivered by Germany on Working Definitions of LAWS / "Definition of Systems under Consideration"', Reaching Critical Will, [online] 9 April, Available from: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Germany.pdf [Accessed 07 March 2024].
Rarm, L. (2023) 'Terror: live', Continuum, 37(3): 422 32. https://doi.org/10.1080/10304312.2023.2255395
Rasmussen, N. (2011) 'Medical Science and the Military: The Allies' Use of Amphetamine during World War II', The Journal of Interdisciplinary History, 42(2): 205–33. https://doi.org/10.1162/JINH_a_00212
Richardson. M, (2020) 'Drone cultures: encounters with everyday militarisms', Continuum 34(6): 858-69. https://doi.org/10.1080/10304312.2020.1842125
Rupka, S., Baggiarini, B. (2018) 'The (non) event of state terror: drones and divine violence', Critical Studies on Terrorism, 11(2): 342-56. https://doi.org/10.1080/17539153.2018.1456735
Rudloff, J.P., Appel, M. (2023) ,Fake News', in M. Appel, F. Hutmacher, C. Mengelkamp, J.P. Stein, S. Weber (eds) Digital ist besser?! Psychologie der Online- und Mobilkommunikation, Berlin und Heidelberg: Springer, pp 217-32. https://doi.org/10.1007/978-3-662-66608-1_15
Sarasin, P. (2006) Anthrax: Bioterror as fact and fantasy, Harvard University Press.
Scharre, P. (2018) Army of None: Autonomous Weapons and the Future of War, New York: W. W. Norton & Company.
Sauer, F. (2016) 'Stopping 'killer robots': why now is the time to ban autonomous weapons systems', Arms Control Today, 46(8).
Schwarz, E. (2018) Killing Machines, Manchester: Manchester University Press.
Seib, P. (2021) Information at War: Journalism, Disinformation, and Modern Warfare, Cambridge, UK: Polity Press.
Shaw, I.G.R. (2016) Predator Empire: Drone Warfare and Full Spectrum Dominance, Minneapolis, London: University of Minnesota Press.

Singer, P.W. (2010) Wired for War: The Robotics Revolution and Conflict in the Twenty-First Century, New York, NY: Penguin Books.
Simonite, T. (2021) '3 Year After the Project Maven Uproar, Google Cozies to the Pentagon', WIRED, [online] 10 November, Available from: https://www.wired.com/story/3-years-maven-uproar-google-warms-pentagon/ [Accessed 12 March 2024].
Stop Killer Robots (n.d.) 'Homepage', Stop Killer Robots [online], Available from: https://www.stopkillerrobots.org/ [Accessed 05 March 2024].
Stoner, R. H. (2009) 'R2D2 with Attitude: The Story of the Phalanx Close-In Weapons, NavWeaps, [online] 30 October, Available from: http://www.navweaps.com/index_tech/tech-103.php [Accessed 07 March 2024].
Suchman, L. (2023) 'The uncontroversial 'thingness' of AI', Big Data & Society, 10(2): 1–5. https://doi.org/10.1177/20539517231206794
The Red Team (2021) 'Discover the Red Team', The Red Team, [online], Available from: https://redteamdefense.org/en/meet-the-red-team [Accessed 12 March 2024].
Toffler, A., Toffler, H. (1993) War and anti-war. Making Sense of Today's Global Chaos, New York, NY: Warner Books.
UK Ministry of Defence/Bundeswehr Office for Defence Planning (2021) Human Augmentation. The Dawn of a New Paradigm, UK Ministry of Defence.
US Department of Defense (2023) 'Dod Directive 3000.09 Autonomy In Weapon Systems, Washington Headquarters Service, [online] 25 January, Available from: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf [Accessed 13 March 2024].
Von Clausewitz, C. (1942) ,Principles of War', Clausewitz Studies, [online], Available from: https://www.clausewitzstudies.org/mobile/principlesofwar.htm [Accessed 12 March 2024].
Williams, A. (2015) 'Defining Autonomy in Systems: Challenges and Solutions', in A. Williams and P. Scharre (eds) Autonomous Systems. Issue for Defence Policymakers, Norfolk, Virginia: NATO Headquarters Supreme Allied Commander.Transformation, pp 27-62.
Worcester, M. (2015) 'Autonomous Warfare – A Revolution in Military Affairs', ISPSW Strategy Series: Focus on Defense and International Security, 340: 1-6.
Work, R.O., Brimley, S. (2014) 20YY: Preparing for war in the robotic age, Center for a New American Century.
Zikusoka, D. (2023) 'How Ukraine's "Uber for Artillery" is Leading the Software War Against Russia', New America, [online] 25 May, Available from: https://www.newamerica.org/future-frontlines/blogs/how-ukraines-uber-for-artillery-is-leading-the-software-war-against-russia/ [Accessed 11 March 2024].

PART III. REASSESSMENT OF THEORETICAL FOUNDATIONS

ARTICLE V. Technology Hypes: Practices, Approaches and Assessments

ARTICLE V

Technology Hypes: Practices, Approaches and Assessments[1]

Jascha Bareis, Maximilian Roßmann, Frédérique Bordignon

Abstract

To date, the study of hype has become a productive but also eclectic field of research. This introduction provides an overview of the core characteristics of technology hype and distinguishes it from other future-oriented concepts. Further, the authors present promising approaches from various disciplines for studying, critiquing, and dealing with hype. The special issue assembles case studies, methodological and theoretical contributions that analyze tech hypes' temporality, agency, and institutional dynamics. It provides insights into how hypes are triggered and fostered, but also how they can be deconstructed and anticipated.

[1] Published 13 December 2023 under CC-BY license in Journal for Technology Assessment in Theory and Practice. Accessed under:https://www.tatup.de/index.php/tatup/article/view/7074/. Content and citation style of the original publication have been adopted.

Introduction

Technology assessment (TA) has been highly productive in discussing the power and problems of technology expectations, futuristic communication, and their overpromising. Situating TA in societal context, ranging from political debates to the attention economy in social media, sheds light not only on the analysis of hype but also on the 'modulation' of visions to reach wider audiences. This may include unheard or neglected voices and arguments in technology development and its critiques, e.g., to reach sustainable development goals (Dierkes et al. 1996; Grunwald 2015; Rip 2006; Schneider et al. 2023).

In contrast to 'vision' or 'expectation', calling technology 'hype' is both descriptive and action-guiding. It suggests a temporal dynamic of attention and confidence in projected technological change – an increase followed by a decrease – and points to the question of inappropriate attitude and reaction, given the context of a debate. At stake are taking poor public policy decisions, misdirecting financial resources, the lack of studying more pressing societal consequences, and, more generally, jeopardizing trust in science (Intemann 2020; Löfstedt 2003).

However, TA has never been alone in developing methods to study and find a response to technology hype. This Special topic in the *Journal for Technology Assessment in Theory and Practice* seeks to highlight the variety of approaches from different disciplines and the internationality of cases. Herewith the issue contributes to a better understanding of temporalities, agency, and institutional dynamics that provoke, fuel and maintain hypes, and provides knowledge to better anticipate, deconstruct and criticize them.

Joint efforts to narrow down the phenomenon: dimensions and characteristics of technology hypes

Rhetorics and the emotional appeal of overpromising language

By means of bold statements, superlatives and exaggerated claims, hypes appeal to emotions to seek attention. Historical analogies to break-throughs or reference to fictional literature serve tech-evangelists to claim proficiency and reliable guidance in uncertain times.

Cherishing narratives of approaching disruptions suggest societal roles and call for requirements to be met, so a specific goal can be achieved (Mische 2014; Van Lente and Rip 1998). People often share technology narratives for the sake of excitement, however, often ignoring how they assemble and change the meaning of arguments, facts, and data, e.g. at a scenario workshop (Roßmann 2021).

Ideally, "imagination under constraints" of beliefs and scientific knowledge allow for societal learning (Kind 2016, p. 3). The simulated experience of technological consequences (by means of illustrative imagery, or stories in place of an argumentation) reaches wider audiences and can help to bridge boundaries between disciplines, publics and institutions (Dierkes et al. 1996; Lösch 2006). However, by means of emotional appeal and dramatization, narrative communication can also bypass the rational assessment of statements (Green and Brock 2000). This emotional celebration of statements is characteristic of hype – and risks turning an informative and appealing story into a sensationalist one.

Social media has further increased this phenomenon. Big tech platforms reinforce outreach and attention to a topic by a system of likes, shares, hails and reposts. Here, research has shown that communication and algorithmic content moderation on platforms supports sensationalism and click-baiting. It is emotional and controversial posts, and especially visual material over factual and descriptive content, that become featured in timelines by users (Gillespie 2018; Gorwa et al. 2020). Such attention-seeking logic on platforms certainly contributes to an environment that nourishes hyping as it elicits emotional appeal and impulsive action over critical reflection.

Given this large influence of language, deliberate and responsible communication about technology requires reflection about potentially conflicting communication aims. It urges us to carefully consider the context, speech positions and audience, when technological novelties are announced.

Temporality and the play with attention spans

The three 'musketeers' of rhetoric – *ethos*, *pathos* and *logos* – are occasionally supplemented by *kairos*, which is the opportune moment for action. Hypes gain their real performative momentum by pointing to vast opportunities that lie ahead, which ask for the right timing if

great potentials shall not be lost. Hence, temporality is a crucial dimension for understanding and negotiating technology expectations.

Popular technology narratives structure salient societal discourses on technology and usually refer to bigger time-spans. The studies of socio-technical imaginaries (Jasanoff and Kim 2009), for example, reveal differences in common understandings of technology projects in society, informing about hopes and concerns in project proposals or policy papers that consider large future trajectories (e.g. see Bareis and Katzenbach 2022 for staging Artificial Intelligence; or Mosco 2005 for the study of the cyberspace metaphor).

Technology hypes, however, radically focus on temporal prominence. Stressing the opportunity costs is a distinctive feature of hypers, who urge followers to act instantly, take risks and think boldly. Thereby hype narrows down remembrances of the past and, likewise, future trajectories to come. While narratives, visions and imaginaries rather mark the cultural background that persists over a longer period of time, technology hypes foreground peak and outlier achievements of tech development. Hereby, they give relevance to certain claims for only a limited period of time. Hypers are the opportunists among future tellers, who ride on the wave of attention and are less interested in the long-term societal consequences of what happens when the wave collapses. Comparing hype cycles therefore studies the attention and popularity of technologies and their claims by means of time-row analysis of publication counts in newspapers or social media, citation counts, or patent applications (Dedehayir and Steinert 2016). One can also draw on discourse analysis or stakeholder interviews, for instance to assess confidence in stock market trends. The representation of a hype cycle according to Gartner Consulting, which follows the evolvement of hype from an attention trigger, over a peak of inflated expectations, to a trough of disillusionment – until state of affairs stabilize in a plateau of productivity, is particularly popular (Linden and Fenn 2003). Though, due to its missing empirical validation, weak theoretical grounding, and instrumental use for claiming future developments with the authority of a seemingly scientific representation, Rip (2006) calls the model a “folk theory” (p. 362).

Time is a crucial factor in the phenomenon of hype – both as a constitutive feature (hypes need the future trajectory in order to gain momentum), and also as an analytical dimension (e.g., when studying the attention span in the building up and waning of a hype).

Impression management and the creation of followership and collaboration

The possibility to learn from imagining futures and to influence how others imagine them with pretense practices, invites various stakeholders for strategic actions and engage in the "politics of expectations" (Beckert 2016, p. 79). Recalling *kairos* above: Observing a trend as hype points to a short window of opportunity to instrumentally exploit the attention for one's own purpose. Especially on social media, the strategic use of certain buzzwords, hashtags and prefixes, like *AI*, *nano-*, *smart-*, or *green-*, helps actors to reach a wider audience, even though actors know that there is little or no shared understanding of the term (Bensaude Vincent 2014). The relationship between leaders and their followers is shaped sustainably by 'impression management' that instills attention and authority in promises about products, applications, or tech-companies. It becomes visible when one follows actors and objects across different sites, revealing differences between 'front region' performances and statements and actions 'backstage' in team meetings or the laboratory (Goffman 1990, p. 69). Technology presentations, such as the release of a new iPhone (Sharma and Grant 2011) or the advertisement of air-taxis (Woznica 2022) strategically highlight and disguise expectations. As 'narrative accelerators' they fuel public discourse and can further bloat an emerging bubble (Goldfarb and Kirsch 2019). In interaction with their own communities of practice, scientists tend to easily reject certain visions but still strategically use these narratives to gain funding or legitimacy from politicians (Selin 2007). Birch (2017), therefore, understands not the expectations of successful technological applications but the expectation to increase the value of research assets, such as networks, laboratory equipment, or topical knowledge, as a major driving force in techno-scientific capitalism. The economic, social, and cultural capital required for (strategic) 'future making' also sheds light on unequal speaker positions to advocate for neglected concerns or more 'profane' and less technocentric visions, such as job opportunities (Sand 2019).

Finally, imagined futures serve as a projective space to coordinate actions (Van Lente and Rip 1998). Sharing problem perspectives and indicating how one would act if a certain scenario unfolds, generates a common ground for individuals or organizations to understand each other and plan with mutual assurances. Two extreme poles can be distinguished that both allow for coordination: either a situation of mutual trust, where stakeholders understand and

rely on each other, or the situation of mistrust, when all statements about the future are perceived as strategic performances resulting from profit or power striving. The study of hypes and overpromising provides insights into popular expectations and their reactions. Mische (2014) suggests developing digital methods to study 'projective grammars' that can further indicate e.g., the perceived openness and attitude of different actors to shape or collaborate in the future.

Although our call for papers drew attention to the fact that digitization of mass media also necessitates a revision of methods for studying imagined futures and that we are particularly interested in computational methods, we received hardly any submissions from this field. In our opinion, TA is a welcoming interdisciplinary niche for experimenting with new methodological approaches. We would therefore call our colleagues to follow up, e.g., with the study of hype language in scientific publications by word lists (Bordignon et al. 2021; Millar et al. 2019; Vinkers et al. 2015), or with the use of metrics of significance (like citation surge or betweenness centrality) to identify emerging trends and potential hypes (Chen 2006; Chen et al. 2012).

Dealing with hype: How and when to intervene?

Actors can be stuck in 'lock-ins' when promises call for action and stakeholders are on the spot to deliver on their bold claims. Such lock-ins hinder organizations to acknowledge 'uncomfortable knowledge' or to share relevant information, which can spur worrisome trajectories based on misguided beliefs (Rayner 2012). Exchanging expectations about potential but unproven harms or benefits of technology is indispensable for reflecting about societal change, though. It is the realm of shared imagined futures that allows for debates, self-reflection and strategic planning about the use and misuse of technology and their societal consequences.

How, though, can we assess when a red line is crossed regarding economic market power and an overheated discursive situation? When do some players gain too much attention and lock society in unwanted path-dependencies? Assessing the discourse on Nanoethics, Nordmann (2007) prominently warned of the looming danger of futuristic 'tunnel visions' that draw all attention and 'ethical resources' away from other, more pressing issues. Also, Vinsel (2021)

understands the criti-hype as an academic business model. Others argue that it may only be right that TA not only analyzes but speaks out for the instrumental use of visions, e.g., to foster democratic values and sustainable development goals (Dierkes et al. 1996; Schneider et al. 2023). Grunwald (2010) argued that enabling public debates about technology in society makes imaginaries available for technology development and can, thus, justify or outweigh the danger of tunnel visions. However, the question remains when and why such an instrumental use of imagined futures becomes inappropriate. Auch (2013) suggests that there is no checklist answer but we can only train our 'virtue of proportionality'. As Dani Shanley illustrates in her TATuP interview (this issue), the history of TA and Responsible Research and Innovation (RRI) also provides some learnings on this.

An even more hands-on treatment of technology hypes would be the building of scenario pathways. Here, policy makers can discuss potential future trajectories and ground lofty discourses with plausibility. This helps them to assess the complexity and ambiguity of future developments and structure messy and contradictory future discussions. The benefits are manifold. Policy makers can escape dominant thought patterns and dive into different epistemic and power positions of actors in society, giving space to silenced and neglected discourses. The biggest benefit of scenarios in the context of hype, though, is to strengthen one's own strategic orientation in the midst of societal crisis, or technology glorification by some attention-seeking actor. The knowledge about different scenarios allows policy makers a strategic-resilient treatment of exuberant promises, encouraging them not to jump on every bandwagon a tech-hype proclaims.

Presentation of the volume

The contributors to this TATuP Special topic have used different methods to respond to our call to deconstruct technological hypes: Some have developed an original analytical framework, others have used interviews and field observations, some have proposed case studies, and finally a few others have also supplemented their study with a quantitative approach.

Roberson et al. examine the dynamics of hype in the field of quantum technology by deconstructing core arguments presented in national strategies. Their analysis considers how

this policy discourse is collaboratively shaped by scientists, politicians, and industry. They challenge current models of hype in science and innovation, mainly the Gartner hype cycle, and propose the 'hype helix', a model that captures the cyclical and iterative nature of hype in research.

Arora and Sarkar endeavor to go beyond hype as a discursive process by redescribing it also as a mnemonic device. They show how tech hype, when applied to emerging technologies like blockchain, can influence the way complex societal problems, such as land rights in India, are (mis-)remembered. Their study highlights the danger of oversimplification and selective presentation of benefits – mainly a solution to corruption and an improvement of land titles management – which overlooks the complexities and nuances of India's land tenure system and the potential negative consequences for marginalized groups.

In his study of exaggerations in debates surrounding social experiments, Neuwinger also finds a tendency among both advocates and critics to overstate benefits and understate risks. This stems from a reductive, tool-based mindset that glosses over complexity by equating social experiments with drug trials, and solely defining impact in causal terms.

Züger et al. demonstrate how the performative nature of expectations has significant implications for actors within the public interest AI field. Their research, employing case studies and interviews, unveils the paradoxical position of actors in public interest initiatives. While they gain support and benefit from the community-building which fosters AI hype, they also maintain a critical stance, acknowledging the risks of unreliable funding and emphasizing the priority of addressing societal needs.

Kari et al. leverage the sociology of expectations perspective to offer valuable insights into the intricate interplay of hype and promises within the domain of nuclear technologies, particularly small modular reactors (SMRs). With the analysis of publication counts and 'hype language' in a Finnish newspaper, they highlight the crucial role of techno-scientific promising in shaping innovation trajectories. They show how the media serves as a key arena where proponents and critics battle over SMRs promises (e.g., cutting carbon emissions and enhancing energy security) leading to SMR topicality, hyping, and eventual deconstruction.

Meunier and Herzog clarify the relationship between a long-term socio-technical imaginary, such as precision medicine, and shorter-term technological hypes, including advancements in omics and AI technologies. They consider that an improvement in the assessment of precision medicine requires a cautious and realistic approach that considers the long-term developments, including previous disappointments, as well as limitations that have hindered the realization of promises being made.

Both Frisch and Gaillard et al. unpack the concept of overpromising and provide new definitions. Frisch sees overpromising as a distinct feature of companies' imagined business futures in response to decarbonization pressure. He suggests that overpromises emerge from contradictions between a company's inevitable profit orientation, the exaggeration and misrepresentation of an organization's estimated potential to restructure itself, and the systemic pressure and bandwagon of performative commitments. Eventually, promoting optimistic narratives about achieving a decarbonized economy can paradoxically hinder climate action by creating a false sense of achievement and delaying necessary measures.

Gaillard et al. explore overpromising as a common feature of scientific discourse, particularly in fields such as nanoscience. In their multidisciplinary approach, combining signaling theory, philosophy of promising, and science studies research on scientific communication, they put forth a conceptualization that facilitates the identification and assessment of overpromises. They emphasize the importance of considering the context of knowledge available when assessing promises and delineating the crucial factors for assessing the plausibility of claims being made.

Some of the case studies that the authors have chosen to present raise issues that ethicists should help to address. But according to Pichl, ethicists can also contribute to hype as she shows in an investigation within the field of stem cells, where therapeutic promises are often used as moral arguments for funding and research-friendly regulation. Pichl's research article clearly demonstrates how this contributes to the hype surrounding stem cell research and its potential applications. To avoid contributing to hype, the article argues, ethicists must critically examine future visions and promises, be aware of their own performative role, and cooperate more closely with disciplines like STS and TA to contextualize analyses within socio-technical dimensions.

We conclude by expressing our gratitude to all the reviewers who contributed to improving the quality of the manuscripts with their constructive comments, and by wishing (with no overpromising) that the readers of this TATuP Special topic will find both inspiration and answers for future work.

References

Auch, Adam (2013): Virtuous argumentation and the challenges of hype. In: Ontario Society for the Study of Argumentation 10: Virtues of Argumentation. Available online at https://core.ac.uk/download/pdf/72768221.pdf, last accessed on 07. 11. 2023.

Bareis, Jascha; Katzenbach, Christian (2022): Talking AI into being. The narratives and imaginaries of national AI strategies and their performative politics. In: Science, Technology, & Human Values 47 (5), pp. 855–881. https://doi.org/10.1177/01622439211030007

Beckert, Jens (2016): Imagined futures. Fictional expectations and capitalist dynamics. Cambridge, MA: Harvard University Press.

Bensaude Vincent, Bernadette (2014): The politics of buzzwords at the interface of technoscience, market and society. The case of 'public engagement in science'. In: Public Understanding of Science 23 (3), pp. 238–253. https://doi.org/10.1177/0963662513515371

Birch, Kean (2017): Rethinking value in the bio-economy. Finance, assetization, and the management of value. In: Science, Technology, & Human Values 42 (3), pp. 460–490. https://doi.org/10.1177/0162243916661633

Bordignon, Frederique; Ermakova, Liana; Noel, Marianne (2021): Over-promotion and caution in abstracts of preprints during the COVID-19 crisis. In: Learned Publishing. Journal of the Association of Learned and Professional Society Publishers 34 (4), pp. 622–636. https://doi.org/10.1002/leap.1411

Chen, Chaomei (2006): CiteSpace II. Detecting and visualizing emerging trends and transient patterns in scientific literature. In: Journal of the American Society for Information Science and Technology 57 (3), pp. 359–377. https://doi.org/10.1002/asi.20317

Chen, Chaomei; Hu, Zhigang; Liu, Shengbo; Tseng, Hung (2012): Emerging trends in regenerative medicine. A scientometric analysis in CiteSpace. In: Expert Opinion on Biological Therapy 12 (5), pp. 593–608. https://doi.org/10.1517/14712598.2012.674507

Dedehayir, Ozgur; Steinert, Martin (2016): The hype cycle model. A review and future directions. In: Technological Forecasting and Social Change 108, pp. 28–41. https://doi.org/10.1016/j.techfore.2016.04.005

Dierkes, Meinolf; Hoffmann, Ute; Marz, Lutz (1996): Visions of technology. Social and institutional factors shaping the development of new technologies. Frankfurt: Campus.

Gillespie, Tarleton (2018): Custodians of the internet. Platforms, content moderation, and the hidden decisions that shape social media. New Haven, CT: Yale University Press. https://doi.org/10.12987/9780300235029

Goffman, Erving (1990): The presentation of self in everyday life. London: Penguin.

Goldfarb, Brent; Kirsch, David (2019): Bubbles and crashes. The boom and bust of technological innovation. Stanford: Stanford University Press. https://doi.org/10.1515/9781503607934

Gorwa, Robert; Binns, Reuben; Katzenbach, Christian (2020): Algorithmic content moderation. Technical and political challenges in the automation of platform governance. In: Big Data & Society 7 (1), p. 205395171989794. https://doi.org/10.1177/2053951719897945

Green, Melanie; Brock, Timothy (2000): The role of transportation in the persuasiveness of public narratives. In: Journal of Personality and Social Psychology 79 (5), pp. 701–721. https://doi.org/10.1037/0022-3514.79.5.701
Grunwald, Armin (2015): Die hermeneutische Erweiterung der Technikfolgenabschätzung. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 24 (2), pp. 65–69. https://doi.org/10.14512/tatup.24.2.65
Grunwald, Armin (2010): From speculative nanoethics to explorative philosophy of nanotechnology. In: NanoEthics 4 (2), pp. 91–101. https://doi.org/10.1007/s11569-010-0088-5
Intemann, Kristen (2020): Understanding the problem of "hype". Exaggeration, values, and trust in science. In: Canadian Journal of Philosophy Cambridge University Press 52 (3), https://doi.org/10.1017/can.2020.45
Jasanoff, Sheila; Kim, Sang-Hyun (2009): Containing the atom. Sociotechnical imaginaries and nuclear power in the United States and South Korea. In: Minerva 47 (2), pp. 119–146. https://doi.org/10.1007/s11024-009-9124-4
Kind, Amy (2016): Imagining under constraints. In: Amy Kind and Peter Kung (eds.): Knowledge through imagination. Oxford: Oxford University Press, pp. 145–159. https://doi.org/10.1093/acprof:oso/9780198716808.003.0007
Linden, Alexander; Fenn, Jackie (2003): Understanding Gartner's hype cycles. Strategic analysis report. Available online at http://ask-force.org/web/Discourse/Linden-HypeCycle-2003.pdf, last accessed on07. 11. 2023.
Löfstedt, Ragnar (2003): Science communication and the swedish acrylamide 'alarm'. In: Journal of Health Communication 8 (5), pp. 407–432. https://doi.org/10.1080/713852123
Lösch, Andreas (2006): Anticipating the futures of nanotechnology. Visionary images as means of communication. In: Technology Analysis & Strategic Management Routledge, 18 (3–4), pp. 393–409. https://doi.org/10.1080/09537320600777168
Millar, Neil; Salager-Meyer, Françoise; Budgell, Brian (2019): "It is important to reinforce the importance of …". 'Hype' in reports of randomized controlled trials. In: English for Specific Purposes 54, pp. 139–151. https://doi.org/10.1016/j.esp.2019.02.004
Mische, Ann (2014): Measuring futures in action. Projective grammars in the Rio+20 debates. In: Theory and Society 43 (3–4), pp. 437–464. https://doi.org/10.1007/s11186-014-9226-3
Mosco, Vincent (2005): The digital sublime. Myth, power, and cyberspace. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/2433.001.0001
Nordmann, Alfred (2007): If and then. A critique of speculative NanoEthics. In: NanoEthics 1 (1), pp. 31–46. https://doi.org/10.1007/s11569-007-0007-6
Rayner, Steve (2012): Uncomfortable knowledge. The social construction of ignorance in science and environmental policy discourses. In: Economy and Society 41 (1), pp. 123–125. https://doi.org/10.1080/03085147.2011.637335
Rip, Arie (2006): Folk theories of nanotechnologists. In: Science as Culture 15 (4), pp. 349–365. https://doi.org/10.1080/09505430601022676
Roßmann, Maximilian (2021): Vision as make-believe. How narratives and models represent sociotechnical futures. In: Journal of Responsible Innovation 8 (1), pp. 70–93. https://doi.org/10.1080/23299460.2020.1853395
Sand, Martin (2019): On "not having a future." In: Futures 107, pp. 98–106.

https://doi.org/10.1016/j.futures.2019.01.002
Schneider, Christoph; Roßmann, Maximilian; Lösch, Andreas; Grunwald, Armin (2023): Transformative vision assessment and 3-D printing futures. A new approach of technology assessment to address grand societal challenges. In: IEEE Transactions on Engineering Management 70 (3), pp. 1089–1098. https://doi.org/10.1109/TEM.2021.3129834
Selin, Cynthia (2007): Expectations and the emergence of nanotechnology. In: Science, Technology, & Human Values 32 (2), pp. 196–220. https://doi.org/10.1177/0162243906296918
Sharma, Abz; Grant, David (2011): Narrative, drama and charismatic leadership: The case of Apple's Steve Jobs. In: Leadership 7 (1), pp. 3–26. https://doi.org/10.1177/1742715010386777
Van Lente, Harro; Rip, Arie (1998): Chapter 7. Expectations in technological developments. An example of prospective structures to be filled in by agency. In: Cornelis Disco and Barend Van der Meulen (eds.): Getting new technologies together. Berlin: De Gruyter, pp. 203–230. https://doi.org/10.1515/9783110810721.203
Vinkers, Christiaan; Tijdink, Joeri; Otte, Willem (2015): Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014. Retrospective analysis. In: British Medical Journal, p. h6467. https://doi.org/10.1136/bmj.h6467
Vinsel, Lee (2021): You're doing it wrong. Notes on criticism and technology hype. In: Medium, 01. 02. 2021. Available online at https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5, last accessed on 07. 11. 2023.
Woznica, Marcel (2022): Stage performances as means for linking sociotechnical imaginaries and projective genres in the discourse around urban air mobility. In: European Journal of Futures Research 10 (12). https://doi.org/10.1186/s40309-022-00198-3